TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG
Die Ressourcenuniversität. Seit 1765.

HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

Bioinformatische Analysen zur Optimierung von Aptamerbibliotheken

Eine Studie zur Aufklärung bindungsrelevanter Charakteristika in Sequenz und Struktur am Beispiel eines Norovirus-Aptamers

Von der Fakultät für Chemie und Physik
der Technischen Universität Bergakademie Freiberg

genehmigte

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt von **M. Sc. Rico Beier**

geboren am 22. März 1988 in Burgstädt

Gutachter: Prof. Dr. rer. nat. habil. Michael Schlömann, TU Bergakademie Freiberg
Prof. Dr. rer. nat. Dirk Labudde, Hochschule Mittweida

Tag der Verleihung: Freiberg, den 29. November 2018

Bibliographische Angaben

Beier, Rico: Bioinformatische Analysen zur Optimierung von Aptamerbibliotheken — Eine Studie zur Aufklärung bindungsrelevanter Charakteristika in Sequenz und Struktur am Beispiel eines Norovirus-Aptamers, Technische Universität Bergakademie Freiberg, Fakultät für Chemie und Physik.

Dissertation, 29. November 2018.

Die Arbeit umfasst insgesamt 256 Seiten, ohne Verzeichnisse und Erklärungen 192 Seiten. Sie beinhaltet 74 Abbildungen, 22 Tabellen, 54 Formeln und 566 Literaturangaben.

Satz: LaTeX.

Referat

Aptamere besitzen als nahezu universelle Binder ein großes Anwendungspotential in Biotechnologie und Medizin; ihre Selektion aus zufälligen Sequenzbibliotheken liefert jedoch nur suboptimale Ergebnisse. Während der Selektion schlagen sich in der Bibliothek Bindungsinformationen nieder, die zur Optimierung des Verfahrens eingesetzt werden können. Die vorliegende Arbeit widmet sich der bioinformatischen Erschließung dieser Informationen. Für die initiale Auswertung der gewonnenen Aptamersequenzdaten erwiesen sich unter Zuhilfenahme von n-Gramm-Deskriptoren und Affinitäten die Regressionsanalyse und auf statistisch-symbolischer Ebene die Mustersuche als wirkungsvolle Verfahren. Für die Aufklärung der Komplexstruktur aus Aptamer und Zielprotein wurde eine Bewertungsfunktion gefunden, die die Identifikation sogenannter nahe-nativer Bindungskonformationen unter den Ergebnissen einer Dockingsimulation erlaubt. Nach dieser methodischen Evaluation erfolgte die Selektion eines Aptamers gegen das Kapsid des humanen Norovirus. Auf Basis der erhobenen Sequenzdaten wurde die Bindegeometrie zwischen Aptamer und Zielprotein durch Anwendung der im Verfahrensprotokoll festgehaltenen Kombination der Analysemethoden aufgeklärt. Das dabei als bindungsrelevant identifizierte Sequenzmotiv kann bei der Erzeugung targetspezifisch optimierter Selektionsbibliotheken als *a priori*-Information einfließen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

Abkürzungsverzeichnis

AIR *Ambiguous Interaction Restraints* . . . 125
AK Autokorrelation . . . 56
AM1 *Austin Model 1* (Semiempirische Methode der Quantenberechnung) . . . 54
AMBER *Assisted Model Building with Energy Refinement* (Kraftfeld) . . . 105
BCNI *Binary Coded Nucleobase Information* (Deskriptoren) . . . 54
BSA *Bovine serum albumin* (dt. 'Bovines Serumalbumin') . . . 147
CASP *Critical Assessment of Techniques for Protein Structure Prediction* (Evaluation computergestützter Methoden der Proteinfaltung) . . . 180
CCD *Charge-coupled device* (dt. 'ladungsgekoppeltes Bauteil') . . . 152
CCI *Charge-Charge Interaction* (Deskriptoren) . . . 54
CDK *Chemistry Development Kit* (Programmierbibliothek) . . . 60
CE *Combinatorial Extension* . . . 183
CHARMM *Chemistry at Harvard Macromolecular Mechanics* (Kraftfeld) . . . 105
CUDA *Compute Unified Device Architecture* (Programmiertechnik zur Ansteuerung des Grafikprozessors) . . . 187
DARS-RNP *Decoys As the Reference State Potential* (Bewertungsfunktion) . . . 103
DECK-RP *Distance- and Environment-dependent, Coarse-grained and Knowledge-based RNA/Protein* (Paarweise Bewertungsfunktion) . . . 104
DFIRE *Distance-scaled, Finite, Ideal-gas Reference* (Bewertungsfunktion) . . . 103
DNA *Deoxyribonucleic Acid* (dt. 'Desoxyribonukleinsäure') . . . 32
dsDNA *Double Stranded DNA* (dt. 'Doppelsträngige DNA') . . . 45
EDTA Ethylendiamintetraessigsäure . . . 46
EMA *Evolution-mimicking algorithm* . . . 123
EM Energieminimierung . . . 186
FG-MD *Fragment-Guided Molecular Dynamics* (Webserver zur Vorhersage von Proteinstrukturen) . . . 183
GE *General Electric* . . . 146
GmbH Gesellschaft mit beschränkter Haftung . . . 146
GRIND *Grid-independent molecular Descriptors* (Deskriptoren) . . . 59
GROMACS *Groningen Machine for Chemical Simulations* (Simulationssoftware) . . . 186
GROMOS *Groningen Molecular Simulation* (Kraftfeld) . . . 105
HADDOCK *High Ambiguity Driven Biomolecular Docking* (Simulationssoftware) . . . 105
IPLF *ISIDA Property-Labelled Fragments* (Deskriptoren) . . . 59
IRMSD *Interface RMSD* (Wurzel aus der mittleren quadratischen Abweichung im Bereich der molekularen Schnittstelle) . . . 110
ISIDA *In Silico Design and Data Analysis* . . . 60

ISR *Intrinsic Specificity Ratio* (dt. 'Intrinsische Spezifität') . . . 107
I-TASSER *Iterative Threading Assembly Refinement* (Webserver zur Vorhersage von Proteinstrukturen) . . . 180
ITScore-PP *Iterative Knowledge-based Scoring Function for Protein-Protein interactions* (Bewertungsfunktion) . . . 104
ITScore-PR *Iterative Knowledge-based Scoring Function for Protein-RNA interactions* (Bewertungsfunktion) . . . 104
JNI *Java Native Interface* (Programmierschnittstelle) . . . 108
JPPF *Java Parallel Processing Framework* . . . 109
KGaA Kommanditgesellschaft auf Aktien . . . 147
KK Kreuzautokorrelation . . . 57
KSV Komplexitäts-Status-Vektor . . . 86
KV Kreuzvalidierung . . . 65
LAT längenabhängige Transformation . . . 56
LED *Light-emitting Diode* (dt. 'Lichtemittierende Diode') . . . 152
LMO *Leave-Multiple-Out* (Datenverteilungsstrategie bei der KV) . . . 79
LNA *Locked Nucleic Acid* (dt. 'Verschlossene Nukleinsäure') . . . 43
logRS logarithmische, relative Promotorstärke . . . 62
LOMETS *Local Meta-Threading-Server* (Webserver zur Strukturvorhersage) . . . 181
LOO *Leave-One-Out* (Datenverteilungsstrategie bei der KV) . . . 79
LUT längenunabhängige Transformation . . . 63
MIF *3D Molecular Interaction Field* (Deskriptoren) . . . 59
MMB *Macromolecule Builder* (Simulationssoftware) . . . 185
MMTSB *Multiscale Modeling Tools for Structural Biology* (Simulationssoftware) . 121
MOPAC *Molecular Orbital Package* (Simulationssoftware) . . . 54
MPI *Message Passing Interface* (Standard für den Nachrichtenaustausch in der parallelen Verarbeitung) . . . 187
mRNA *Messenger RNA* (dt. 'Boten-RNA') . . . 33
NBI *Non-Bonded Interaction* (Deskriptoren) . . . 54
NGS *Next Generation Sequencing* (Neue Sequenzierungsmethode) . . . 47
NPIDB *Nucleic Acids-Protein Interaction Database* . . . 108
NP *Nondeterministic Polynomial time* (dt. 'Nichtdeterministisch polynomielle Zeit') . . . 88
OPLS *Optimized Potentials for Liquid Simulations* (Kraftfeld) . . . 105
ORF *Open Reading Frame* (dt. 'Offener Leserahmen') . . . 141
PCR *Polymerase Chain Reaction* (dt. 'Polymerase-Kettenreaktion') . . . 45
PDB *Protein Data Bank* . . . 100
PLS *Partial Least Squares* (Mathematisches Verfahren) . . . 66
PME *Particle Mesh Ewald* (Energetisches Berechnungsschema) . . . 187
PNA *Peptide Nucleic Acid* (dt. 'Peptid-Nukleinsäure') . . . 43
PSPM *Position-Specific Presense Matrix* (Repräsentation variabler Muster) . . . 84

PSSM *Position-Specific Scoring Matrix* (Repräsentation variabler Muster) 83
QSAR *Quantitative Structure-Activity Relationship* (dt. 'Quantitative Struktur-Wirkungs-Beziehung') 59
QUASI-RNP *Quasi-chemical Potential* (Bewertungsfunktion) 103
RHDV *Rabbit Hemorrhagic Disease Virus* (dt. 'Chinaseuche-Virus') 181
RMSD *Root-Mean-Square Derivation* (dt. 'Wurzel aus der mittleren quadratischen Abweichung') 110
RMSE *Root-Mean-Square Error* (dt. 'Wurzel aus dem mittleren quadratischen Fehler') 65
RNA *Ribonucleic Acid* (dt. 'Ribonukleinsäure') 32
RPS *RNA Polymerase Site* (Promotorregion) 61
rRNA Ribosomale RNA 124
RT-PCR *Reverse Transcription PCR* 46
RT-qPCR *Reverse Transcription Quantitative PCR* 143
SBS *Sequencing by Synthesis* (dt. 'Sequenzierung durch Synthetisierung') . . 147
SDS Natriumlaurylsulfat 46
SELEX *Systematic Evolution of Ligands by Exponential enrichment* (Experimentelles Verfahren zur Gewinnung von Aptameren) 43
SGBP *Scores of Generalized Base Properties* (Deskriptoren) 53
SMF *Substructural Molecular Fragments* (Deskriptoren) 59
SPA-PN *Specificity and Affinity of the Protein-Nucleic acid Interactions* (Bewertungsfunktion) 104
SPR *Surface plasmon resonance* (dt. 'Oberflächenplasmonenresonanz') 152
ssDNA *Single Stranded DNA* (dt. 'Einzelsträngige DNA') 39
SSI Sekundärstrukturinformation 59
ThreaDom *Threading-based Protein Domain Prediction* (Webserver zur Vorhersage von Proteinstrukturen) 181
TM *Template modeling* 182
UPGMA *Unweighted Pair Group Method with Arithmetic Mean* (Methoden zum Clustern von Daten) 160
Weka *Waikato Environment for Knowledge Analysis* (Softwarepaket) 67
XML *Extensible Markup Language* 109
XRD *X-Ray Diffraction* (dt. 'Röntgendiffraktion', Experimentelle Strukturaufklärung) 108

Nasci non potest
Singuli vi
Traditum tale
Opus Rici.

— RG / RB

Vorwort

Ein Werk wie dieses kann nicht aus der Kraft eines Einzelnen entstehen. Wenn auch ein Einzelner die Feder führt, so führen doch Viele diesen Einen. Ich möchte daher an dieser Stelle die Chance ergreifen und all denen danken, die direkt oder indirekt am Zustandekommen meiner Dissertation beteiligt waren. Wahrscheinlich werde ich dabei nicht alle namentlich benennen können, seht es mir bitte nach. Im Inneren weiß ich um Jeden.

In erster Linie möchte ich mich bei meinen beiden Betreuern Prof. Dr. Michael Schlömann und Prof. Dr. Dirk Labudde für die Möglichkeit bedanken, dass ich diese Arbeit am Institut für Biowissenschaften der TU Bergakademie Freiberg sowie im Rahmen der *Bioinformatics Group* Mittweida durchführen konnte. Die Visionen der ersten Tage wurden für meine Arbeit zum Keim der thematischen Gestaltung, der von der fachlichen Begleitung genährt über die letzten Jahre zu einer wissenschaftlichen Arbeit heranwachsen konnte.

Von großer Wichtigkeit war dabei auch die experimentelle Expertise, die mir durch die TU Dresden zur Verfügung gestellt wurde. Sowohl für die Planung und Durchführung der Aptamerselektion als auch für die experimentelle Validierung der Aptamerbindung gilt mein Dank den Mitarbeitern der Bioverfahrenstechnik des Instituts für Naturstofftechnik. Ich danke ferner der *Deep Sequencing Group* des Biotechnologischen Zentrums, die über die Sequenzierung der Selektionsbibliotheken die Schnittstelle zwischen Experiment und bioinformatischer Analyse herstellte. Ferner möchte ich Claudia Pahlke und Dr. Elke Boschke für die konstruktive Arbeit an unserer gemeinsamen Publikation danken.

Unserem Administrator Marcel Scheuche möchte ich für die Bereitstellung der Hardware- und Softwareinfrastruktur danken, durch die es möglich wurde, meine Berechnungen parallel durchzuführen. Eine große Bedeutung hatte zudem die Einbettung in die wissenschaftliche Gemeinschaft vor Ort, die mir in der *Bioinformatics Group* Mittweida mit zahlreichen, konstruktiven Diskussionen und einem steten Gedankenaustausch geschaffen wurde. Ich danke allen Mitgliedern der Arbeitsgruppe. Für sein besonders kritisches Auge und die gegenseitige Hilfe bei der organisatorischen Bewältigung der Promotion möchte ich mich bei Florian Heinke bedanken. Schließlich danke ich meinem Zimmernachbarn Stefan Schildbach dafür, dass seine Gelassenheit zu mancher Stunde mein Gemüt beruhigte.

Gerade in der Schlussphase sind mir einige Menschen bei der Korrektur der Arbeit besonders intensiv beiseite gestanden. Für ihre hilfreichen Korrekturen möchte ich Florian Heinke, Anne-Marie Pflugbeil und Prof. Dr. Dirk Labudde danken. Mein besonderer Dank gilt in diesem Kontext jedoch meiner Freundin Rebekka Grunwald und meinem Kollegen Tommy Bergmann, die sich

beide in großer Anstrengung, Geduld und Detailtiefe durch die gesamte Dissertation gearbeitet haben. Mit diesem nicht selbstverständlichen Engagement haben Sie maßgeblich zur Qualität der schriftlichen Arbeit beigetragen.

Zuletzt möchte ich meiner Familie, meinen Freunden und besonders meiner Freundin für all die herzliche Unterstützung und Motivation danken, die sie mir in der Zeit meiner Promotion zukommen ließen. Sie haben besonders in den letzten beiden Jahren stets Verständnis für meine gedankliche und physische Abwesenheit aufgebracht und mir damit einen unschätzbaren Freiraum als Reserve geschaffen, auf den ich die Last verteilen konnte.

Das Promotionsvorhaben wurde im Zeitraum von 2012 bis 2014
aus Mitteln des Europäischen Sozialfonds und des Freistaates Sachsen
im Rahmen eines Promotionsstipendiums finanziell gefördert.

1 Thematische Einleitung

Aptamere bilden eine wichtige Klasse künstlich erzeugter, funktioneller Nukleinsäuren, die eine chemische Bindung mit einem breiten Spektrum von Zielmolekülen eingehen können. Dieses Spektrum reicht von kleinen organischen Molekülen bis hin zu komplexen Makrostrukturen, wobei den Proteinen als wichtigsten Funktionsträgern aller Lebensprozesse eine besondere Bedeutung zugemessen wird. Aptamere treten damit in Konkurrenz zu den bereits etablierten Antikörpern und erschließen sich aufgrund zahlreicher Vorteile ein großes Anwendungspotential in den Bereichen der biotechnologischen Industrie und der Medizin. Für ein gegebenes Zielmolekül können hochaffine Aptamere mit dem Verfahren SELEX aus einer zufällig synthetisierten Bibliothek von Aptamerkandidaten gewonnen werden, ohne dass targetspezifische Informationen *a priori* bekannt sein müssen. Hierzu wird in einem mehrstufigen Selektionsprozess ein künstlicher Selektionsdruck geschaffen, der innerhalb der Bibliothek zu einer sukzessiven Anreicherung derjenigen Aptamerkandidaten führt, die eine hohe Affinität zum Zielmolekül aufweisen. Die Affinität geht dabei in der Regel nicht von der gesamten Aptameroberfläche aus, sondern wird von spezifischen Bindebereichen des Aptamers vermittelt. Für die experimentelle Durchführbarkeit der Aptamerselektion unterliegt die synthetisierte Bibliothek jedoch einer Größenbeschränkung und kann daher nicht die Gesamtheit aller möglichen Aptamerkandidaten abdecken. Als großer Nachteil des Verfahrens muss daher festgestellt werden, dass es sich bei den selektierten Aptameren mit großer Sicherheit um Vertreter mit suboptimaler Affinität zum Zielmolekül handelt. Durch die stattfindende Ausdifferenzierung zwischen bindenden und nichtbindenden Aptamerkandidaten schlagen sich jedoch während der Selektion in der Bibliothek Informationen zu den Bindepräferenzen des Zielmoleküls nieder. Diese Informationen können zur Erzeugung einer targetspezifisch optimierten Bibliothek und damit zum Finden leistungsfähigerer Aptamere verwendet werden, müssen zunächst aber aus der Bibliothek extrahiert werden.

Ziel dieser Arbeit ist es daher, ein bioinformatisches Analyseprotokoll zu etablieren, mit dessen Hilfe die in der Bibliothek verborgenen Informationen zu den Bindepräferenzen extrahiert, beschrieben und für ihre Optimierung nutzbar gemacht werden können. Aus dieser Zielvorgabe leiten sich für die Arbeit zwei inhaltliche Schwerpunkte ab. Der erste Schwerpunkt umfasst die Suche und Evaluierung zielstellungsrelevanter, bioinformatischer Analysemethoden. Auf Basis der Sequenz- und Sekundärstrukturdaten einer Selektionsbibliothek sollen dabei numerische Deskriptoren eingesetzt werden, um in Bezug zur Affinität der Aptamerkandidaten Sequenzbereiche zu bestimmen, die die Bindung realisieren. Auf statistisch-symbolischer Ebene soll dies durch die Betrachtung überrepräsentierter Teilsequenzen ohne ein solches biologisches Referenzkriterium im Rahmen einer Mustersuche ergänzt werden. Zur Aufklärung der tatsächlichen Bindegeometrie zwischen Aptamer und Zielprotein ist ferner die Bewertung der Kandidatenkomplexe einer Docking-Simulation notwendig. Die hierfür genutzte paarweise Bewertungsfunktion soll in der Lage sein, ohne Vorhandensein eines Referenzkomplexes einzuschätzen, ob die Bindegeometrien der simulierten Komplexe der unbekannten natürlichen Geometrie ähnlich sind. Der zweite Schwerpunkt liegt auf der kombinierten Anwendung der Analyseverfahren am konkreten Beispiel. Zu diesem Zweck wird der gesamte Arbeitsprozess mitsamt der

Aptamerselektion gegen ein Kapsidprotein des humanen Norovirus und der darauffolgenden Datenerhebung abgebildet. Neben der Aufklärung der Bindungsgeometrie zwischen Aptamer und Zielprotein sollen die Analysen Bindecharakteristika liefern, die für die Optimierung der Selektionsbibliothek eingesetzt werden können.

1.1 Hypothesen und Fragestellungen

Für die sukzessive Erarbeitung der gewählten Zielstellung werden aus den zwei inhaltlichen Schwerpunkten insgesamt vier elementare Hypothesen und konkretisierende Fragestellungen abgeleitet. Diese beziehen sich auf die vorgestellten Kategorien bioinformatischer Analysemethoden und deren kombinierte Anwendung am konkreten Beispiel. Die Hypothesen bilden damit einen Leitfaden durch die Arbeit, der mit der Auswahl der Analysemethoden beginnt und mit deren kombinierter Anwendung endet.

Hypothese 1 Die physikochemische Prägung numerischer Deskriptoren ist bei der Charakterisierung von Aptamersequenzen entscheidend für die Güte der abgeleiteten Beschreibung.

Fragestellung 1.1 Welche Deskriptoren kommen allgemein für die Beschreibung von Nukleinsäuresequenzen infrage?

Fragestellung 1.2 Welche Arten von Information fließen in die Berechnung dieser Deskriptoren ein?

Fragestellung 1.3 Wie kann die Güte der Deskriptoren an einem Datensatz verlässlich evaluiert werden?

Fragestellung 1.4 Welche informationellen Komponenten wirken sich auf die Güte der Deskriptoren maßgeblich aus?

Hypothese 2 Auch unter der Anforderung von Variabilität ist die Mustersuche in großen Datensätzen von Aptamersequenzen mit begrenzter Rechenkapazität durchführbar.

Fragestellung 2.1 Welche Verknüpfung aus Datenstruktur und Algorithmus eignet sich besonders für die Mustersuche in großen Sequenzdatensätzen?

Fragestellung 2.2 Wie kann die Variabilität biologischer Sequenzen zweckmäßig für den Einsatz in Mustern abgebildet werden?

Fragestellung 2.3 Wie kann das aufgestellte Konzept von Variabilität in die Datenstruktur und den Algorithmus integriert werden?

Fragestellung 2.4 Welche Faktoren limitieren die algorithmische Umsetzung der Mustersuche?

Hypothese 3 Paarweise Bewertungsfunktionen können die interatomaren Kontakte zwischen Proteinen und Aptameren selbst dann zur Beurteilung eines simulierten Komplexes nutzen, wenn keine Referenzstruktur zum Vergleich vorliegt.

Fragestellung 3.1 Welche Beschreibungskonzepte eignen sich für die Bewertung von Protein-Nukleinsäure-Komplexen?

Fragestellung 3.2 Sind diese auch ohne Verfügbarkeit eines Referenzkomplexes in der Lage, nahe- von nicht-nativen Kandidatenstrukturen zu unterscheiden?

Fragestellung 3.3 Können die Ergebnisse einer solchen Bewertung auch für den fallübergreifenden Vergleich von Strukturen genutzt werden?

Hypothese 4 Die bioinformatische Analyse erhobener Sequenzdaten erlaubt die Aufklärung der Bindekonstellation zwischen Aptamer und Zielprotein sowie die Erkennung bindungsrelevanter Sequenz- und Strukturmerkmale, die zum Entwurf einer optimierten Selektionsbibliothek eingesetzt werden können.

Fragestellung 4.1 Erlaubt das etablierte Verfahren SELEX die erfolgreiche Selektion eines Aptamers gegen das große Kapsidprotein `VP1` des humanen Norovirus Genotyp `GII.4`?

Fragestellung 4.2 Lassen sich die eingeführten bioinformatischen Analyseverfahren auf die veränderte Datenbasis übertragen, die durch die erhobenen Sequenzdaten der Aptamerselektion gegeben ist?

Fragestellung 4.3 Können die Analyseverfahren in ihrer gemeinsamen Anwendung die Bindekonstellation zwischen Aptamer und Zielprotein aufklären?

Fragestellung 4.4 Lassen sich für das gefundene Aptamer auf dieser Basis Merkmale der Sequenz und Struktur ableiten, die relevant für die beobachtete Bindung sind und sich für die Optimierung der Selektionsbibliothek eignen?

1.2 Aufbau der Arbeit

Als Handreichung zur weiteren Orientierung wird der Aufbau dieser Arbeit in einer kapitelweisen Darstellung in den folgenden Absätzen beschrieben. Die im Laufe dieser Arbeit entstandenen Publikationen werden dabei in den zugehörigen Kontext eingeordnet. Zum visuellen Überblick dient die Gliederung in Abbildung 1.1. Mit dem Ende dieses einleitenden Kapitels sind sowohl Motivation als auch Zielstellung und Hypothesen der Arbeit klar herausgearbeitet. Im 2. Kapitel werden anschließend wichtige Grundlagen vermittelt, die zum Verständnis der darauffolgenden Kapitel essentiell sind. Diese umfassen den Aufbau von Proteinen und Nukleinsäuren sowie einen kurzen Abriss zum Thema der Protein-Nukleinsäure-Komplexe. Auf dieser Basis werden

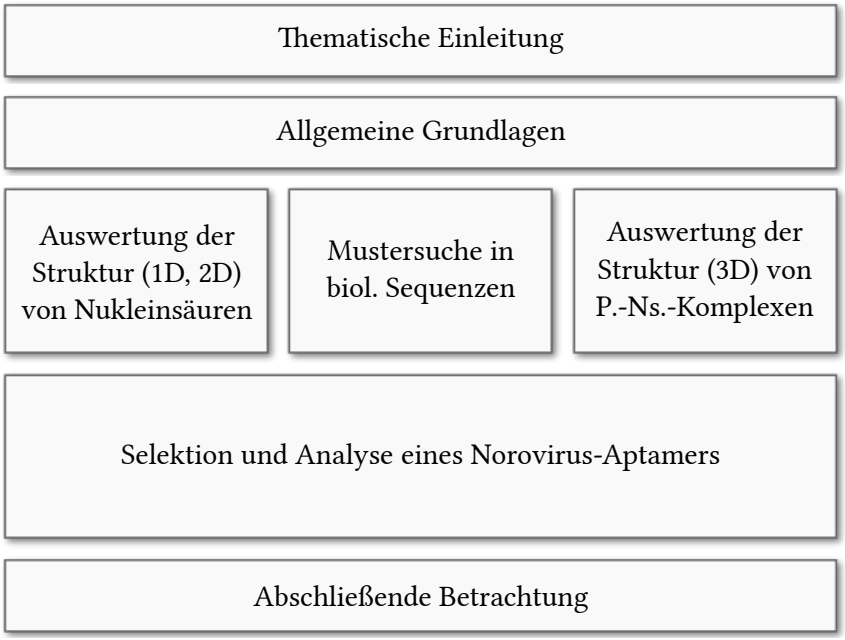


Abb. 1.1: Kapitelweise Gliederung dieser Arbeit

Aptamere als wichtige Vertreter der funktionellen Nukleinsäuren mit ihren relevanten Anwendungsfeldern eingeführt. Das Kapitel schließt mit der Vorstellung des Verfahrens SELEX, das zur Generierung targetspezifischer Aptamere gemeinhin verwendet wird.

Der erste inhaltliche Hauptteil beschäftigt sich in den drei Kapiteln 3 bis 5 mit der Evaluation von zielstellungsrelevanten bioinformatischen Analysemethoden. Der Fokus des 3. Kapitels liegt dabei auf der Beschreibung von Nukleinsäuresequenzen durch numerische Deskriptoren. Unter Einbezug physikochemischer Eigenschaften, sowie der Primär- und Sekundärstrukturinformationen werden relevante Beschreibungsmodelle vorgestellt und auf einen konkreten Datensatz funktionaler Nukleinsäuren angewendet, um die Relevanz der aus ihnen abgeleiteten Deskriptoren miteinander zu vergleichen. Die Ergebnisse des Kapitels wurden zum großen Teil bereits in einer begleitenden Publikation [1] veröffentlicht. Als eigenständige thematische Ergänzung widmet sich das 4. Kapitel der Auswertung biologischer Sequenzdaten auf rein statistisch-symbolischer Ebene. Die Entwicklung eines Suchverfahrens für überrepräsentierte Sequenzmuster zielt dabei auf die Anwendbarkeit für große Sequenzdatensätze und die Integration sequenzieller Variabilität. Eine vorläufige Variante des Suchverfahrens wurde in weniger formalisierter Art bereits in einer begleitenden Publikation [2] veröffentlicht. Schließlich adressiert das 5. Kapitel die Bewertung von simulierten Nukleinsäure-Protein-Komplexen anhand ihrer Bindungsgeometrie. Nach der initialen Systematisierung existierender Beschreibungskonzepte folgt die detaillierte Vorstellung ausgewählter Vertreter. Im kritischen Vergleich werden an Beispielkomplexen wichtige Eigenschaften der gewählten Bewertungsfunktionen herausgearbeitet und gegenübergestellt. Dies erlaubt eine Einschätzung darüber, ob die Bewertungsfunktionen auch ohne Vorhandensein einer geeigneten Referenzstruktur in der Lage sind, simulierte Komplexstrukturen als nahe- und nicht-nativ zu klassifizieren. In jedem der drei Kapitel wird dabei jeweils eine methodische Empfehlung herausgearbeitet.

Im zweiten inhaltlichen Hauptteil kommen die empfohlenen Analysemethoden zur konkreten Anwendung. Das 6. Kapitel beginnt dazu mit der detaillierten Vorstellung des Norovirus als Zielstruktur für die Aptamerselektion und beschreibt anschließend die experimentellen Details

dieser Selektion sowie eine erste Beurteilung des Selektionserfolges. Für die gewinnbringende Kombination der Analysemethoden sorgt das eigens dafür eingeführte bioinformatische Analyseprotokoll. Über eine stete Rückkopplung von Teilergebnissen gewährleistet es einen nahezu verlustfreien Informationsfluss zwischen den Verfahren, der eine gegenseitige Evaluierung erlaubt und damit die Plausibilität der Ergebnisse weiter erhöht. Durch Simulation und Analyse wird schließlich die tatsächliche Bindegeometrie des Komplexes aus Aptamer und Zielprotein aufgeklärt und ein bindungsrelevantes Sequenzmuster abgeleitet, welches für die Optimierung der Selektionsbibliothek eingesetzt werden kann. Das Kapitel beschließt mit einem Ausblick auf Konzepte der gezielten Generierung targetspezifischer Bibliotheken. Die Selektion des Aptamers und eine erste Beurteilung des Selektionserfolges wurden bereits in einer begleitenden Publikation [3] veröffentlicht.

Die abschließende Betrachtung des 7. Kapitels nimmt rückblickend Bezug auf die einführend aufgestellten Hypothesen. Nach der prägnanten Beantwortung der zugehörigen Fragestellungen werden die Hypothesen dazu jeweils einzeln verifiziert oder falsifiziert. Die Arbeit endet mit einem kurzen Ausblick.

2 Allgemeine Grundlagen

Zusätzlich zu den themenbezogenen Einführungen der folgenden Kapitel werden vorbereitend einige allgemeine Grundlagen vermittelt, die zum Verständnis der Arbeit notwendig sind. Diese umfassen die Bedeutung und den Aufbau von Proteinen, Nukleinsäuren und Protein-Nukleinsäure-Komplexen, auf die sich alle im Verlauf der Arbeit vorgestellten Beschreibungskonzepte direkt oder indirekt gründen. Die anschließende Vorstellung der Aptamere als spezielle Klasse der funktionellen Nukleinsäuren unterstreicht deren Entwicklungspotential in zahlreichen Anwendungsgebieten. Es schließt sich eine kurze Einführung in das Selektionsverfahren an, welches zur Gewinnung hochaffiner Aptamere für spezielle Zielmoleküle genutzt wird. Als zentrales Verfahren stellt es einen wichtigen Ansatzpunkt für Optimierungen in der Aptamergewinnung dar, der in dieser Arbeit thematisiert wird.

2.1 Protein-Nukleinsäure-Komplexe

Für das Verständnis dieser Arbeit ist eine grundlegende Kenntnis über Proteine, Nukleinsäuren und die aus ihnen gebildeten Komplexe wichtig. Dieser Abschnitt widmet sich daher der einführenden Vorstellung dieser beiden Klassen von Makromolekülen mit Fokus auf deren Bedeutung, Aufbau und physikochemischen Eigenschaften. Abschließend folgt eine kurze Betrachtung der intermolekularen Interaktionen, welche die Bildung von Protein-Nukleinsäure-Komplexen ermöglichen.

2.1.1 Aufbau von Proteinen

Proteine sind kettenförmig aus Aminosäuren aufgebaute Makromoleküle mit essenzieller Bedeutung in der Biologie. Neben den intrinsisch ungeordneten Vertretern werden Proteine prinzipiell drei Klassen unterschieden [4]. Zur ersten dieser Klassen gehören die gut wasserlöslichen, globulären Proteine, die eine kompakte Tertiärstruktur mit tendenziell hydrophobem Kern und weitestgehend hydrophiler Oberfläche ausbilden [5; 6]. Mit ihrem breiten funktionellen Spektrum und der hohen Mobilität in wässriger Umgebung sind globuläre Proteine an nahezu allen biologischen Prozessen des Lebens beteiligt. Sie sind beispielsweise in der Lage, chemische und biologische Reaktionen als Enzyme zu katalysieren [7] und als Inhibitoren zu hindern [8]. Als Bestandteil von Glykoproteinhormonen dienen sie vereinzelt auch der Informationsübertragung [9]. Ihre Fähigkeit der molekularen Erkennung ermöglicht nicht nur die Grundfunktion des Immunsystems [10], sondern lässt sich auch im Rahmen einer technischen Anwendung als Biosensor für die Detektion zunutze machen [11]. Schließlich soll erwähnt werden, dass globuläre Proteine auch toxische Wirkungen entfalten können [12]. Die zweite Gruppe sind die Membranproteine, die durch ihre Einbettung in die nicht-polare Umgebung biologischer Membranen eine großteils hydrophobe Oberfläche aufweisen [13]. Aufgrund der Schwierigkeit, geeignete Kristallisationsbedingungen zu finden, ist ihre strukturelle Aufklärungsquote sehr gering [14]. In der Membran dienen die Proteine vorrangig dem aktiven und passiven Transport

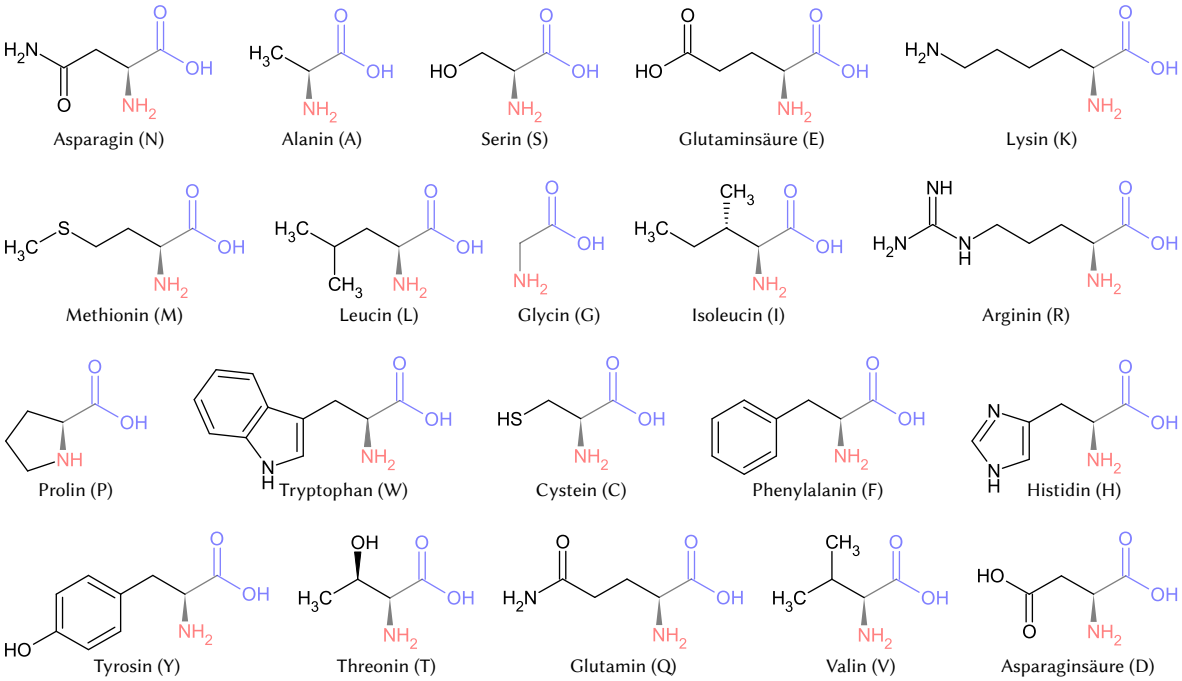


Abb. 2.1: Übersicht der 20 kanonischen, proteinogenen Aminosäuren. Die Carboxyl- (blau) und Aminogruppen (rot) der Hauptkette sind farblich hervorgehoben. Um die wechselnden Seitenketten deutlicher abzuheben, wurden die unveränderlichen Anteile der Hauptketten zudem schwächer koloriert dargestellt.

von Ionen, Wasser und anderen kleinen Molekülen [15–18], sowie der Erkennung von Botenstoffen und physiologischen Veränderungen der Umgebung [19]. Die dritte Gruppe umfasst die strukturell vergleichsweise armen, fibrillären Proteine. Sie sind in der Regel langkettig oder in großen Ringen aufgebaut [20]. Ihre unlösliche, faserartige Struktur dient zur Formgebung und zum Halt von Geweben und anderen biologischen Makrostrukturen [21–23], in denen sie besonders in ihrer Kombination von Bedeutung sind [24].

Aminosäuren als Grundbausteine der Proteine Aminosäuren sind organische Verbindungen, welche über ein zentrales C^{α}-Atom vier spezifische Substituenten vereinen. Dies sind das α-Proton, die Carboxylgruppe, die Aminogruppe und eine variierende Seitenkette. Aus der Vielzahl der existierenden Aminosäuren kommen in natürlichen Proteinen lediglich die 20 kanonischen Vertreter aus Abbildung 2.1 vor. Außer bei Glycin, dessen Seitenkette als einzelnes Wasserstoffatom Symmetrie herstellt, sind die Aminosäuren asymmetrisch und besitzen damit chirale Eigenschaften. Von den zwei existierenden, spiegelbildlichen Konfigurationen sind in natürlichen Proteinen nur die L-Enantiomere zu finden. Die physikochemischen Eigenschaften einer Aminosäure werden maßgeblich durch die variierende Seitenkette bestimmt und beziehen sich hauptsächlich auf deren Größe, das Vorhandensein aromatischer Ringe und das Ladungsverhalten. Mit dem Protonierungsstatus der Amino- und Carboxylgruppen, sowohl der direkten Substituenten als auch innerhalb der Seitenkette, ist die Gesamtladung einer Aminosäure abhängig vom pH-Wert der Umgebung. Es erfolgt eine Unterscheidung in unpolare und polare Aminosäuren, wobei letztere unter physiologischen Bedingungen weiter in geladene und ungeladene Vertreter eingeteilt werden. Die Polarität und Ladungsverteilung hängt dabei eng mit der Wasserlöslichkeit der Aminosäure zusammen [25, S. 150-153; 26, S. 593-598]. Die physikochemischen Eigenschaften der Aminosäuren werden im Venn-Diagramm von Abbildung 2.2 systematisch zugeordnet.

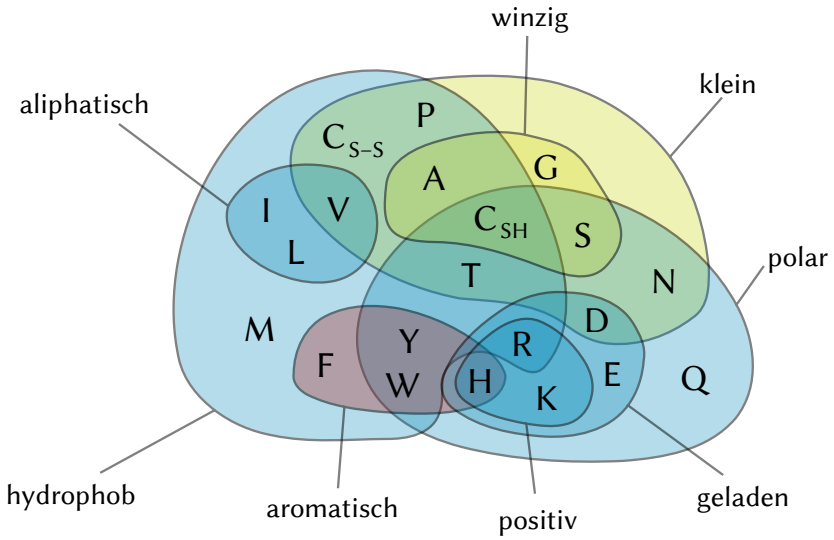


Abb. 2.2: Darstellung der physikochemischen Eigenschaften der 20 kanonischen Aminosäuren im Venn-Diagramm. Die Eigenschaften sind entsprechend einer groben Ordnung eingefärbt (größenbezogen gelb, ladungsbezogen blau, sonstiges rot). Abbildung gestaltet nach [25].

Verknüpfung der Aminosäuren zu Proteinen Die chemische Verbindung der Aminosäuren zu einem Polypeptid oder Protein erfolgt unter Wasserabspaltung in der Polypeptidsynthese, die aufgrund der energetischen Bilanz nicht spontan stattfindet, sondern von Ribosomen katalysiert wird. Dieser Umstand erlaubt die zielgerichtete Synthese anhand einer genetischen Vorlage durch Transkription und Translation. Die Peptidbindung zwischen der Carboxylgruppe der einen und der Aminogruppe der nächsten Aminosäure bildet einen partiellen Doppelbindungscharakter aus, der die Rotation ω um die Peptidbindung stark einschränkt. Damit reduzieren sich die Freiheitsgrade, die den Verlauf des Proteinrückgrates beschreiben, auf die zwei Dieder-Winkel Ψ und Φ um die C^{α}-Atome. Die tatsächliche Bewegungsfreiheit hängt dabei stark von der Beschaffenheit der jeweiligen Seitenketten ab. Stehen sich die Seitenketten auf der gleichen Seite gegenüber, so wird von einer cis-, ansonsten von einer trans-Peptidbindung gesprochen. Bedingt durch die Größe der Seitenketten und die dadurch entstehenden sterischen Kollisionen ist die cis-Variante jedoch wesentlich seltener anzutreffen [25, S. 154-157]. Abbildung 2.3 gibt eine visuelle Darstellung für die Geometrie der Peptidbindung. In der Gesamtheit formt sich durch die Peptidbindungen ein Kohlenstoff-Stickstoff-Rückgrat aus, an welchem entlang die Seitenketten angeordnet sind und somit ein spezifisches physikochemisches Profil bilden.

Neben dieser rückgratbildenden Peptidbindung können kovalente Disulfidbindungen auftreten, wenn sich zwei Cystein-Einheiten geeignet gegenüberstehen. Die geringe Stärke dieser Bindung erlaubt ihre Auflösung jedoch bereits mit milden Reduktionsmitteln [26, S. 607]. Neben den kovalenten Bindungen treten in einer Proteinstruktur nichtkovalente Wechselwirkungen auf. Diese lassen sich in spezifische Vertreter wie Wasserstoffbrücken, Ionenbindungen und aromatische Stapeleffekte sowie unspezifische Vertreter wie die van-der-Waals-Wechselwirkungen einteilen. Sie gehen zu einem großen Teil von den Seitenketten der Aminosäuren aus und führen zu einer stabilen dreidimensionalen Faltung des Proteins [27, S. 165-166].

Sekundärstruktur Durch Wasserstoffbrückenbindungen vermittelt bilden sich lokale, wiederkehrende Strukturmuster in den Proteinen aus, die Sekundärstrukturen genannt werden. Die häufigsten Vertreter sind die rechtsgängigen α-Helices und die β-Faltblätter. Neben diesen exis-

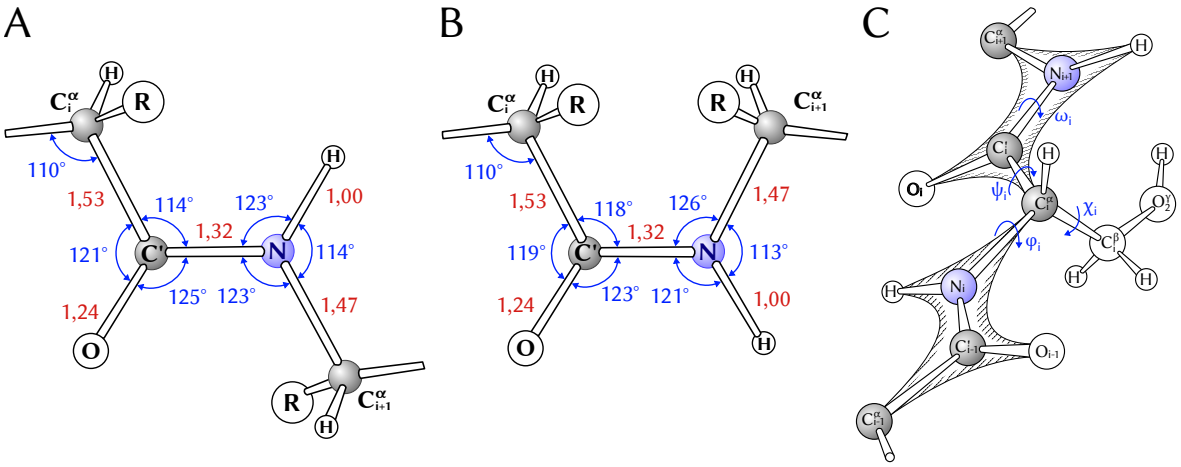


Abb. 2.3: Übersicht über die Geometrie der Peptidbindung. Neben den beiden Bindungsvarianten trans (häufig vorkommend, A) und cis (selten vorkommend, B) wird die Beweglichkeit der Peptidbindung über relevante Drehwinkel (C) dargestellt. Zur besseren Orientierung sind Längenmaße in rot und Winkelmaße in blau hervorgehoben. Die Atome des Peptidrückgrates sind ferner volumetrisch schattiert und entsprechend ihres chemischen Elements koloriert. Gestaltet nach [25, S. 155].

tieren weitere, seltene Vertreter mit energetisch weniger günstigen Konstellationen. Dies sind beispielsweise die rechtsgängigen π- und 3_{10}-Helices, welche meist in sehr kurzer Form vorkommen, sich als letzte Windung an eine α-Helix anschließen oder als strukturelle Unterbrechung in einer α-Helix wirken. Darüber hinaus existieren die linksgängigen α_L- und Kollagenhelices, wobei letztere erst durch die Bildung einer übergeordneten, rechtsgängigen Kollagentripelhelix stabil wird. Bereiche, die keine derartigen wiederkehrenden Sekundärstrukturen enthalten, werden als *Random Coil* bezeichnet [25, S. 158-163; 27, S. 160-163; 28, S. 68; 29; 30].

In einer typischen α-Helix sind jeweils $3{,}6$ Aminosäuren pro Windung rechtsgängig umeinander angeordnet, sodass es zur räumlichen Nachbarschaft von Aminosäuren mit einem sequenziellen Abstand von 4 kommt. Diese Struktur wird durch Wasserstoffbrücken stabilisiert, die sich zwischen dem Amidproton der einen und dem Carbonylsauerstoff der benachbarten Aminosäure ausbilden und aufgrund des Windungsverhältnisses leicht gegenüber der Helixachse geneigt sind. Die Beteiligung aller Peptidgruppen resultiert durch die große Anzahl von Bindungen in einer sehr hohen Stabilität der Helix. Die äußere Lagerung der Seitenketten ermöglicht eine enge Kernformation der Helix mit zusätzlichen, stabilisierenden van-der-Waals-Kontakten im Inneren [25, S. 158-163; 27, S. 160-163].

Unter den β-Faltblättern wird zwischen planaren und nicht-planaren Varianten unterschieden. Im planaren Faltblatt liegen mindestens zwei Aminosäureketten derart im Raum, dass sich zwischen ihren Amidprotonen und Carbonylsauerstoffen seitwärts Wasserstoffbrücken ausbilden. Jeder Strang bildet hierbei namensgebend für das aus ihnen entstehende Faltblatt eine regelmäßig und abwechselnd gefaltete Grundform. Die Stränge eines planaren Faltblatts liegen in einer Ebene, aus der die Seitenketten in der Regel abwechselnd unten und oben nahezu senkrecht herausstoßen. Viel häufiger trifft man jedoch auf nicht-planare Faltblätter, bei denen jeder Strang in sich eine langgezogene, linksgängige Drehung erfährt. Die innere Rotation der Stränge erfordert zur Ausbildung der Wasserstoffbrücken nun auch eine Neigung der Stränge zueinander von etwa $25°$. Abhängig davon, ob die Stränge wie beim parallelen Faltblatt gleichläufig oder wie beim anti-parallelen Faltblatt gegenläufig angeordnet sind, ergeben sich unterschiedliche Bindungsmuster. In größeren Faltblättern sind zudem Kombinationen dieser beiden Orientierungen möglich. Durch die Beteiligung aller nicht am Rand liegenden Aminosäu-

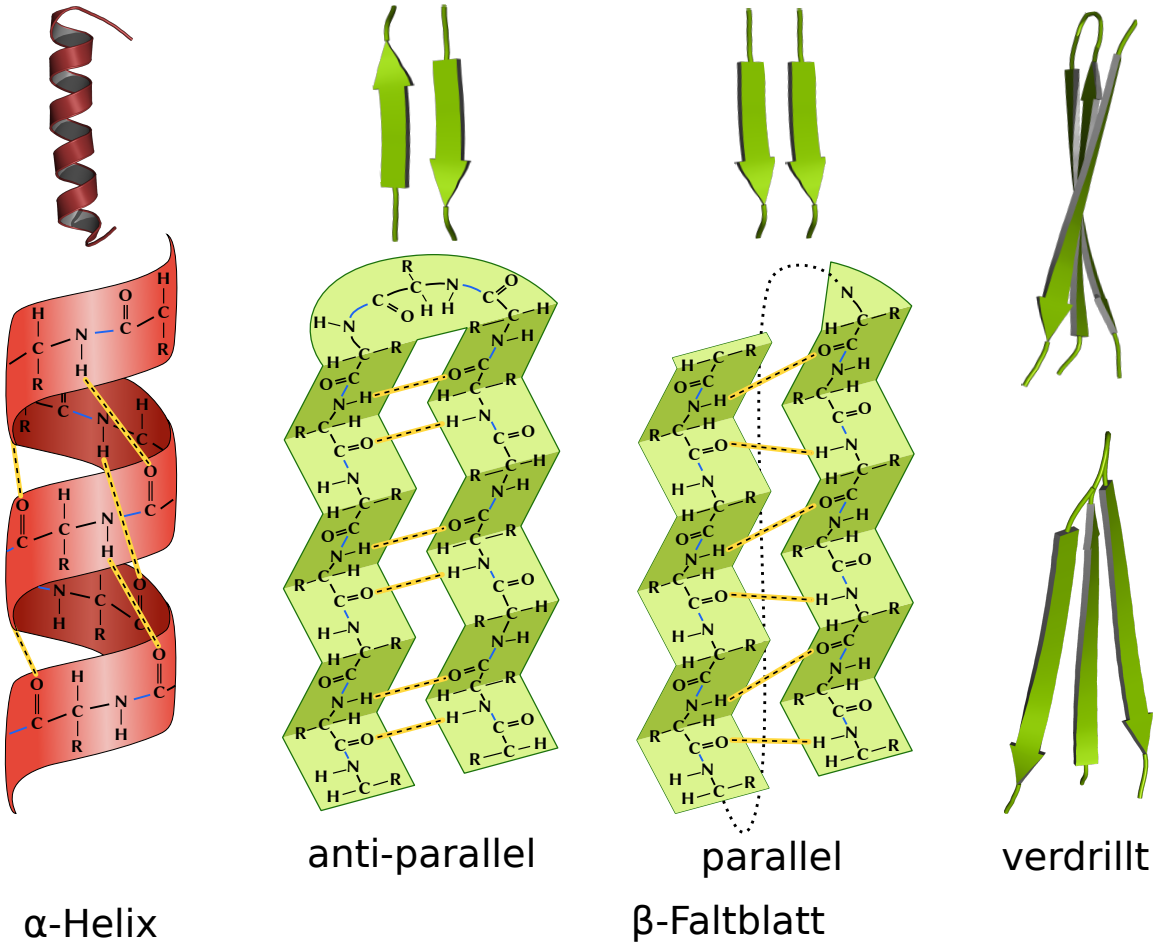


Abb. 2.4: Als wichtigste Sekundärstrukturelemente von Proteinen gelten die α-Helices (rot) und β-Faltblätter (grün), welche in zwei Formen dargestellt sind. Im oberen Bereich der Abbildung befinden sie die in der molekularen Visualisierung üblichen Prinzipdarstellungen, die ohne molekulare und geometrische Details die Lage und den Verlauf der Sekundärstrukturelemente anzeigen. Für die α-Helix und die planaren Varianten des parallelen und anti-parallelen β-Faltblattes werden die molekularen Details im unteren Bereich der Abbildung vereinfacht auf die geometrische Struktur aufgetragen. Zur Orientierung sind Peptidbindungen dabei blau gefärbt und Wasserstoffbrückenbindungen zwischen den Windungen und Faltungen auf gelber Hinterlegung schwarz gestrichelt. Auf die Detailabbildung des verdrillten, nicht-parallelen Faltblattes wurde der Übersicht halber verzichtet. Stattdessen wird dieses aus zwei Perspektiven mit Fokus auf die Neigung der Stränge (oben) und die intrinsische Windung (unten) gezeigt. Teile aus [31].

ren sind β-Faltblätter ebenfalls sehr stabile Sekundärstrukturen [25, S. 158-163; 27, S. 160-163; 26, S. 625-627]. Eine schematische Darstellung der Sekundärstrukturen Helix und Faltblatt findet sich in Abbildung 2.4.

Die vorgestellten Sekundärstrukturelemente zeichnen sich in ihrem Verlauf durch eine sehr ähnliche Orientierung der aufeinanderfolgenden Peptideinheiten aus. Über die Auftragung der Dieder-Winkel Ψ und Φ ergeben sich daher im Ramachandran-Plot der Abbildung 2.5 charakteristische Bereiche für die jeweiligen Sekundärstrukturelemente, die im Rahmen der Validierung und Klassifizierung eingesetzt werden können [25, S. 158-163]. Abhängig von der Größe und den physikochemischen Eigenschaften ihrer Seitenketten unterscheiden sich die Aminosäuren in ihrer Eignung zur Beteiligung an den vorgestellten Strukturen zum Teil deutlich. In einer groben Einteilung werden geeignete Aminosäuren als Bildner und sehr ungeeignete als Brecher der jeweiligen Sekundärstruktur bezeichnet. Eine vollständige und eindeutige Klassifikation ist auf dieser trivialen Ebene jedoch nicht möglich.

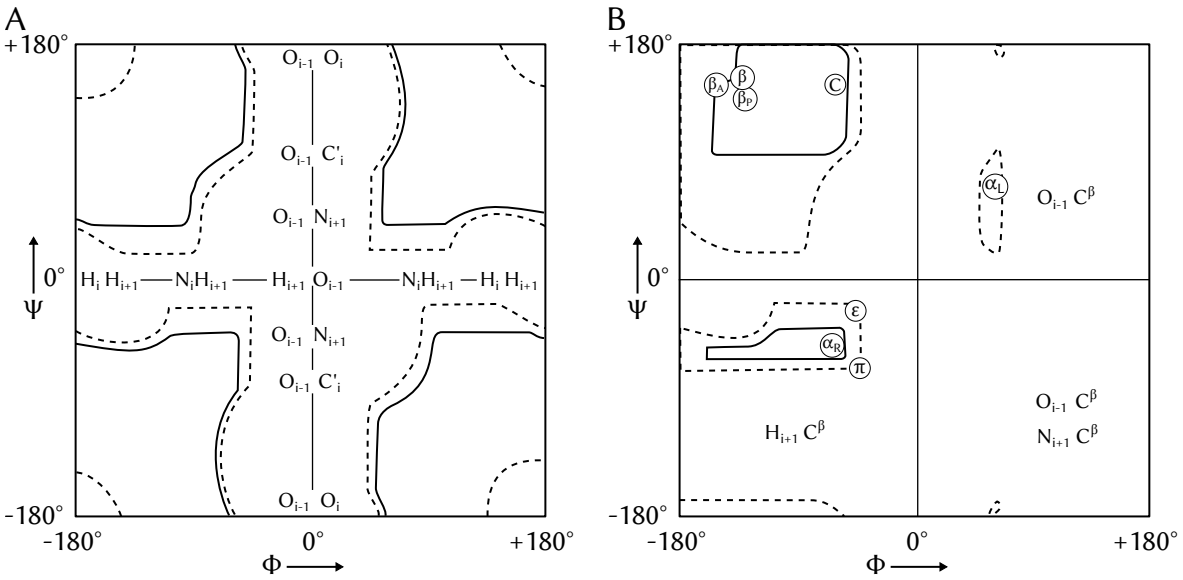


Abb. 2.5: Die zwei Ramachandran-Plots geben die erlaubten Bereiche der Dieder-Winkel Ψ und Φ für einige Protein-Sekundärstrukturen wieder, die mit einem harten Kugelmodell bestimmt wurden. Die gestrichelte Linie gibt dabei die normalen Fälle an, während die durchgezogenen Linien die im Extremfall möglichen Bereiche etwas erweitert darstellt. Ferner sind Atome eingezeichnet, die sich in der jeweiligen Region zu nahe kommen. Die Mobilität der Bindung unterscheidet sich aufgrund der fehlenden Seitenkette markant zwischen Glycin (A) und allen weiteren Aminosäuren (B). Nach [25, S. 156].

Dreidimensionale Proteinstruktur Die Tertiärstruktur beschreibt die räumliche Anordnung einer einzelnen Polypeptidkette, die von den bereits eingeführten kovalenten und nichtkovalenten Interaktionen stabilisiert wird. Obwohl auch Interaktionen mit dem Kohlenstoff-Stickstoff-Rückgrat möglich sind, bildet die sequenzielle Abfolge der Aminosäuren mit ihren spezifischen physikochemischen Eigenschaften und funktionellen Gruppen die Basis für die Ausprägung der Tertiärstruktur. Es bilden sich funktionelle Zentren und Bindungstaschen aus, die für die Funktion der meisten Proteine verantwortlich sind. Da innerhalb der Tertiärstruktur häufig Teilstrukturen beobachtet werden, die voneinander unabhängig in der Lage sind, eine stabile räumliche Anordnung auszubilden, wurde für diese Teilstrukturen der Begriff der Domäne als grundlegende Einheit der Tertiärstruktur geprägt. Häufig ist mit der eigenständigen Struktur einer Domäne auch eine spezifische Funktion assoziiert. Durch die Prädominanz der beiden Sekundärstrukturelemente α-Helix und die β-Faltblatt ist eine Einteilung der Domänen in die Faltungsklassen α, β, $\alpha + \beta$ und α/β üblich. Während einfache und zumeist kleine Proteine aus genau einer solchen Domäne bestehen, gibt es zahlreiche komplexe Proteine, die sich modular aus mehreren Domänen zusammensetzten. Setzt sich ein Protein aus mehreren Polypeptidketten zusammen, so bilden diese jeweils eigene Tertiärstrukturen aus und formen anschließend durch nicht-kovalente Bindung eine gemeinsame Quartärstruktur. Die Untereinheiten können dabei wie im Falle der Homo-Multimere gleichartig oder wie bei den Hetero-Multimeren verschiedenartig sein [25, S. 169-174; 27, S.164].

2.1.2 Aufbau von Nukleinsäuren

Ähnlich wie Proteine sind auch Nukleinsäuren kettenförmig aufgebaute Makromoleküle mit einer sehr hohen Bedeutung in der Biologie. Sie werden auf Basis ihrer Grundbausteine, der Nukleotide, in DNA und RNA unterschieden. Ihre wohl bekannteste Funktion ist die Speicherung

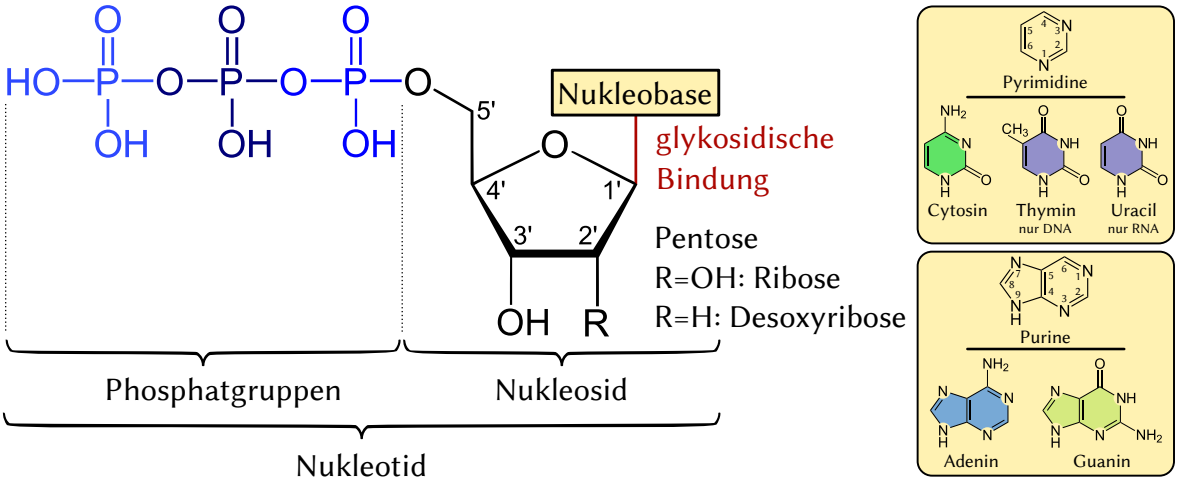


Abb. 2.6: Schematischer Aufbau eines Nukleotids aus Nukleobase, Pentose und Phosphatgruppen. Die fünf möglichen Nukleobasen stammen aus den beiden Gruppen der Pyrimidine und Purine (gelb hinterlegt). Sie werden über eine glykosidische Bindung (rot) mit dem Pentosemolekül zum Nukleosid verbunden. Durch das Anfügen von bis zu drei Phosphatgruppen entsteht schließlich daraus das Nukleotid.

und Übermittlung der genetischen Informationen in Form der DNA und *Messenger RNA* (mRNA). Es existieren jedoch auch vereinzelte Nukleinsäuren mit katalytischer Aktivität, die sogenannten Ribozyme. Sie sind jedoch in den Ribosomen auch an der Translation der mRNA zu Proteinen beteiligt. Darüber hinaus sind Nukleinsäuren in vielfältige Weise in die Regulation der Genexpression involviert. Neben regulierenden Bereichen auf den Chromosomen existieren Formen der RNA, die posttranskriptional und dabei teilweise kontextsensitiv agieren. Mithilfe des Prinzips der RNA-Interferenz wird zellintern beispielsweise die Expression von Genen und die Replikation von Viren durch die gezielte Blockierung funktional wichtiger mRNA mit kurzen RNA-Fragmenten komplementärer Sequenz gestoppt. Im therapeutischen Umfeld können derartige *Small Interfering RNAs* bereits gezielt synthetisiert und zur bewussten Einflussnahme auf die Genexpression eingesetzt werden. Da Nukleinsäuren auch mit anderen Biomolekülen chemische Bindungen eingehen können, eröffnet sich ein weiterer Verwendungszweig, der in Form der künstlich erzeugten Aptamere im nächsten Unterkapitel konkretisiert wird [32].

Nukleotide als Grundbausteine Wie Abbildung 2.6 darstellt, besteht ein Nukleotid grundlegend aus drei Bestandteilen. Der erste Bestandteil ist die Nukleobase. Die fünf kanonischen Nukleobasen lassen sich ihrer Struktur nach in zwei Klassen einteilen. Zur Klasse der Pyrimidine gehören Cytosin, Thymin (nur bei DNA) und Uracil (nur bei RNA). Sie teilen das monozyklische Pyrimidin-Grundgerüst mit Sauerstoff an Position 2 und unterscheiden sich durch die Besetzung der Positionen 4 und 5. Adenin und Guanin gehören zur Klasse der Purine. Sie haben damit das bizyklische Purin-Grundgerüst gemeinsam und unterscheiden sich an den Positionen 2 und 6. Die Nukleobasen weisen damit individuell verschiedene physikochemische Eigenschafen auf. Der zweite Bestandteil ist das D-Enantiomer eines chiralen Pentosederivats. Während bei RNA Ribose zum Einsatz kommt, findet bei DNA die reduzierte Form Desoxyribose Anwendung. Die Nukleobase ist mit dem C_1-Atom dieser Pentose über eine glykosidische Bindung zum sogenannten Nukleosid verbunden. Durch bis zu dreifache Veresterung mit Phosphatgruppen entsteht schließlich das Nukleotid als Grundbaustein der Nukleinsäuren. Durch die zahlreichen polaren Gruppen sind die Nukleotide gut wasserlöslich. Die Verbindung der Nukleotide zu kettenförmigen Polymeren erfolgt über die Ausbildung von Phosphodiesterbindungen. Der bereits

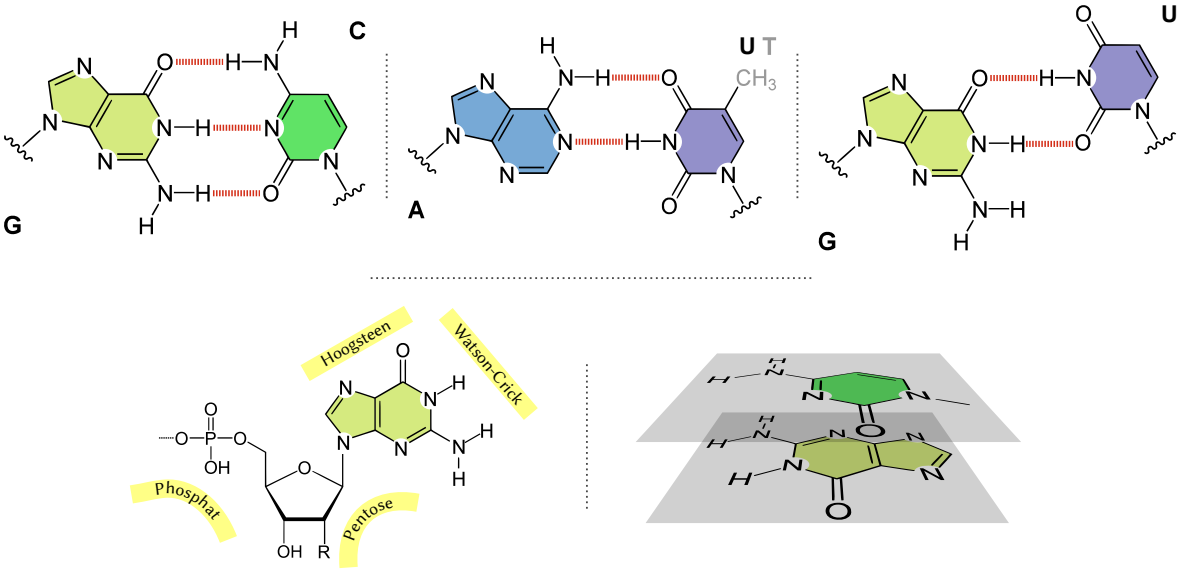


Abb. 2.7: Strukturbildende Effekte bei Nukleinsäuren. In der oberen Zeile befinden sich die bekannten Basenpaarungen nach Watson und Crick (G:C links, A:T/U mittig), sowie die energetisch ungünstigere *Wobble*-Paarung (G:U rechts). In der unteren Zeile folgt eine Übersicht der möglichen Bindestellen eines Nukleotids (links) sowie eine Prinzipdarstellung nahezu planar gestapelter Nukleobasen (rechts).

vorhandene Phosphorsäurerest am 5′-Sauerstoff des einen Nukleotids geht dabei eine zweite Esterbindung mit der 3′-Hydroxylgruppe des folgenden Nukleotids ein [26, S. 639-665; 32, S. 77-80]. Durch die Verbindung hinreichend vieler Nukleotide bildet sich die Nukleinsäure mit einem Zucker-Phosphat-Rückgrat, an dem sequenziell die Nukleobasen verankert sind.

Strukturgebende Faktoren Verantwortlich für die Ausformung von Sekundärstrukturen sind bei Nukleinsäuren die Fähigkeiten der Nukleobasen, über nichtkovalente Wechselwirkungen sogenannte Basenpaarungen und Stapelinteraktionen miteinander einzugehen. Abbildung 2.7 unterstützt die folgenden Erläuterungen. Zwischen den Nukleobasen Adenin und Uracil oder Thymin können zwei Wasserstoffbrücken gebildet werden, während zwischen Cytosin und Guanin drei solche Verbindungen möglich sind. Da bei diesen sogenannten Watson-Crick-Paaren die gleiche Bindungsgeometrie entsteht, kann eine Abfolge dieser Paare auch bei wechselnder Zusammensetzung eine regelmäßige, helikale Struktur ausbilden. Eine Ausnahme bildet hier das *Wobble*-Paar zwischen Guanin und Uracil, das für die Mehrdeutigkeit in der Codon-Anticodon-Wechselwirkung verantwortlich ist, ansonsten jedoch selten vorkommt. Zusätzlich zu den Watson-Crick-Paaren existieren noch zahlreiche weitere Bindungsmodi zwischen Nukleobasen, wie beispielsweise die reversen Watson-Crick-Paare mit umgekehrter Ausrichtung der Bindeflächen. Zudem existieren auch Konstellationen mit anderen Bindebereichen des Nukleotids wie dem ebenfalls nukleobasenspezifischen Hoogsteen-Bereich und dem unspezifisch interagierenden Zucker-Phosphat-Rückgrat. Der energetische Gewinn der Nichtstandard-Paarungen für die Gesamtbindung ist jedoch sehr gering, da mit dem umgebenden Lösungsmittel Wasser ähnlich gute Konstellationen für Wasserstoffbrücken möglich sind. Außerdem eignen sich diese Geometrien nicht für die Bildung von helikalen Strukturen. Sie tragen daher nur mäßig zur energetischen Stabilität der Struktur bei, sondern primär zur spezifischen Formgebung. Besonders in helikalen Bereichen leisten die Stapelwechselwirkungen aufeinanderfolgender Nukleobasen einen großen Beitrag zur Stabilität der Nukleinsäurestruktur. Grund hierfür ist Ausbildung eines transienten, partiellen Dipols zwischen den planar angeordneten,

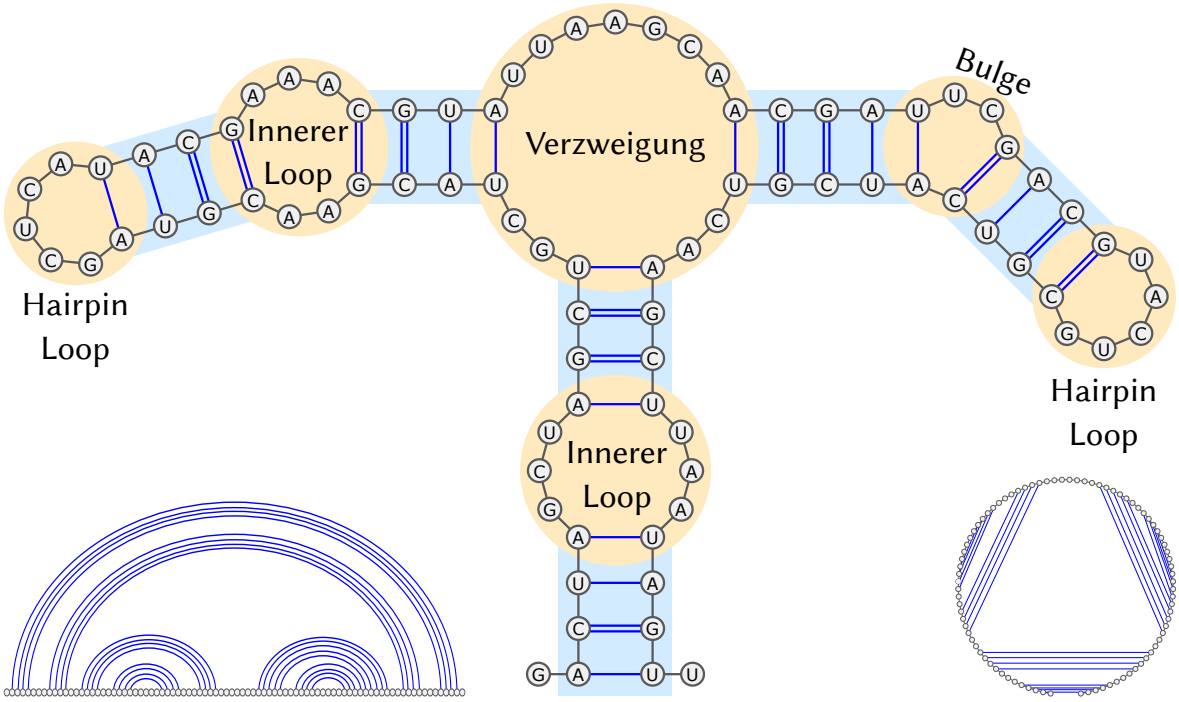


Abb. 2.8: Beispielhafte Übersicht über die helikalen (blau) und unterbrechenden (beige) Sekundärstrukturelemente von Nukleinsäuren. Zur Darstellung wird eine konstruierte Nukleinsäure mit der Sequenz `GACUAGCUAGCUGCUACGAACGUAGCUCAUACGAAACGUAUUAAGCAACGAUUCGACGUACUGCGUCAUCGUCAAGCUUAAUAGUU` verwendet, deren Sekundärstruktur in der üblichen *Dot-Bracket*-Notation durch die Zeichenkette `.((((...((((..((((..((((.....))))...))))).......((((..((((.....))))))))))..))))...))))` spezifiziert ist. Die Ausbildung der helikalen *Stem*-Bereiche wird von den spezifischen Basenpaarungen bestimmt. Die beiden *Hairpin Loops* links und rechts außen sowie der symmetrische innere *Loop* im unteren Bereich beeinflussen die Helixorientierung nicht maßgeblich. Der unsymmetrische inneren *Loop* im oberen, linken Bereich und der *Bulge* auf der rechten Seite führen zu einem Abknicken der Struktur, während die zentrale Verzweigung die drei strukturellen Kernbereiche miteinander verbindet und damit maßgeblich verantwortlich für die Komplexität der Sekundärstruktur ist. Im unteren Bereich der Abbildung werden zwei weitere Visualisierungsformen für die gleiche Sekundärstruktur gezeigt (links linear, rechts zyklisch) [34].

aromatischen Ringsystemen benachbarter Nukleobasen. Tendenziell bilden Guanin und Cytosin stärkere Stapelwechselwirkungen aus als Adenosin, Thymidin und Uridin. Zusätzlich kann auch der hydrophobe Effekt in Nukleinsäuren beobachtet werden, wo er in der Regel die innere Lage der hydrophoben Flächen begünstigt [32, S.80-83; 25, S. 6; 33].

Sekundärstruktur Die Ausbildung von Sekundärstrukturen hängt bei Nukleinsäuren eng mit der Komplementarität der Nukleobasenfolge zusammen. Ideal komplementäre Nukleotidstränge sind in der Lage, durchgehender Bereiche von Watson-Crick-Basenpaarungen auszubilden, die als helikale *Stems* bezeichnet werden und in drei grundlegenden Formen existieren. Form B gilt mit zehn Basenpaaren pro Windung als Normalform einer DNA-Helix. Form A stellt eine kompaktere Variante dar, die mit stark geneigten Nukleobasen alle elf Nukleotide eine Windung formt. RNA bildet durch das sterische Potential der zusätzlichen Hydroxygruppe der Ribose ebenfalls eine Helix der Form A aus. Neben diesen beiden rechtsgängigen Helices existiert die seltene, linksgängige und eher zackig angeordnete Form Z [25, S. 9; 32, S. 82-87].

Bereiche ungepaarter Nukleobasen unterbrechen derartige Helices und formen weitere Sekundärstrukturelemente, wie in Abbildung 2.8 dargestellt. Zwischen zwei Helices gelegene Nukleotide, die an keiner Paarbindung teilnehmen, werden bei einseitigen Vorkommen als *Bul-*

ges und bei beidseitigen als innere *Loops* bezeichnet. Unsymmetrische Varianten dieser beiden Sekundärstrukturelemente haben durch Biegung oder Abknicken großen Einfluss auf den Verlauf der weiterführenden Helix. Liegt ein ungepaarter *Loop*-Bereich am Ende einer Helix, so wird er als *Hairpin Loop* gesondert betrachtet. Zur Überbrückung der Helix sind mindestens vier Nukleobasen notwendig, die strukturelle Stabilität der *Hairpins* nimmt jedoch mit zunehmender *Loop*-Größe ab. Interagieren im weiteren Verlauf der Nukleinsäurekette freie komplementäre Nukleobasen mit denen eines solchen *Hairpin Loops*, bilden sich sogenannte Pseudoknoten. Zur strukturellen Vielfalt der Nukleinsäuren tragen schließlich Verzweigungsstellen bei, die verschiedene Helixverläufe miteinander verbinden [25, S. 8-11; 35; 32, S. 108-109].

Für die einfache digitale Verarbeitung werden Sekundärstrukturen in der sogenannten *Dot-Bracket*-Notation dargestellt, die ungepaarte Nukleobasen durch einen Punkt und Basenpaare durch passende Klammerpaare repräsentiert. Für die visuelle Aufarbeitung eignen sich lineare und zirkuläre Anordnungen der Nukleobasen, in denen Paarungen durch in der Regel überschneidungsfrei verlaufende Linien miteinander verbunden werden. Zur Einschätzung der strukturellen Ausformung sind diese Darstellungen jedoch nicht geeignet. Zu diesem Zweck kann die Sekundärstruktur als Prinzipdarstellung gezeichnet werden, welche Verzweigungen und Knicke wiedergibt [34]. Die genannten Darstellungsformen sind in Abbildung 2.8 exemplarisch gezeigt.

Tertiärstruktur Auf der Basis der Restriktionen, die durch die Basenpaarungen der Sekundärstruktur vorgegeben werden, beschreibt die Tertiärstruktur schließlich die räumliche Anordnung einer Nukleinsäurekette. Diese weist eine wesentlich höhere Komplexität als die Sekundärstruktur auf und wird durch zusätzliche Stapelinteraktionen und Wasserstoffbrücken der Nukleobasen nicht aufeinanderfolgender Nukleotide der räumlichen Umgebung sowie die von Watson-Crick verschiedenen Modi der Basenpaarung stabilisiert. Durch die polaren Gruppen der Nukleotide können divalente Metallionen ferner Einfluss auf die Tertiärstruktur nehmen. Im Rahmen der Tertiärstrukturausbildung kommt es vereinzelt auch zur Interaktion von mehr als zwei Nukleobasen oder zu energetisch bedingten Abweichungen von den Standardkonfigurationen der Sekundärstrukturelemente wie dem Herausklappen einer Nukleobase aus einer helikalen oder unterbrechenden Struktur. Die Tertiärstruktur ist dabei nicht in allen Fällen statisch, sondern kann auch alternative Konformationen umfassen, die abhängig von den physiologischen Umgebungsbedingungen wechseln [25, S. 12-15; 35; 32, S. 83, 110–111].

2.1.3 Interaktionen zwischen Proteinen und Nukleinsäuren

Die Interaktionen zwischen Proteinen und Nukleinsäuren sind in unterschiedlichen biologischen Kontexten von großer Relevanz. Sie ermöglichen beispielsweise die Genexpression, bestimmen die Wirkungsfähigkeit natürlicher funktioneller Nukleinsäuren und sind Grundlage für die Spezifität und Affinität der Bindung künstlicher Nukleinsäuren an ihr Zielmolekül. Die Interaktionen stützen sich dabei auf eine Reihe nicht-kovalenter Wechselwirkungen. So können zwischen Aminosäure und Nukleotid sowohl direkt als auch über größere Distanz durch Wasser vermittelte Wasserstoffbrücken ausgebildet werden und van-der-Waals-Wechselwirkungen auftreten. Abhängig von den physikochemischen Eigenschaften der beteiligten funktionellen Gruppen kann es ferner zu elektrostatischen Wechselwirkungen zwischen geladenen Gruppen, einem Stapeleffekt von aromatischen Ringen sowie im Zusammenspiel mehrerer Residuen auch zu einem Einfluss des hydrophoben Effekts kommen [36–40]. Die räumliche Struktur der beiden

Bindepartner schränkt über die Zugänglichkeit der einzelnen funktionellen Gruppen und deren relative Positionierung die verfügbaren Kontakte ein. Kleinere strukturelle Anpassungen sind jedoch trotz dieser Einschränkung bei energetisch günstiger Lage möglich [36; 41; 42].

Spezifität der Interaktionen Die Interaktionen zwischen Proteinen und Nukleinsäuren lassen sich neben ihrem physikalischen Typ hinsichtlich ihrer sequenziellen Spezifität einteilen. Spezifisch sind dabei jene Interaktionen, die sich beim Nukleotid auf die wechselnde Nukleobase und bei der Aminosäure auf die variierende Seitenkette beziehen. Die Definition erlaubt das Auftreten teilweise spezifischer Interaktionen. Da weder das Stickstoff-Kohlenstoff-Rückgrat bei Proteinen noch das Zucker-Phosphat-Rückgrat bei Nukleinsäuren entlang des makromolekularen Stranges eine sequenzabhängige Unterscheidungskraft besitzt, gelten die Interaktionen mit diesen Bestandteilen als unspezifisch. Sie leisten einen wichtigen, direkten Beitrag zur Stabilisierung der Komplexstruktur und über die stabilisierte Passform der Bindungspartner auch einen kleinen, indirekten Beitrag zu deren Spezifität.

Eine Studie an 129 Komplexen aus Proteinen und DNA-Doppelhelices zeigt die große Bedeutung der unspezifischen Interaktionen auf Seiten der Nukleinsäuren. Sowohl bei den direkten als auch bei den etwa gleich häufig vorkommenden vermittelten Wasserstoffbrücken entfielen etwa zwei drittel der erfassten Kontakte auf das Nukleinsäurerückgrat, wobei die Kontakte zu den Phosphoreinheiten klar dominierten. Unter den Nukleobasen zeigte sich eine positive Tendenz für die Purine während unter den Aminosäuren polare und positiv geladene Residuen in den Interaktionen bevorzugt wurden. Die wesentlich stärker vertretenen van-der-Waals-Kontakte zeigten mit rund drei vierteln einen noch höheren Anteil von unspezifischen Bindungen, der jedoch wesentlich schwächer von der Phosphateinheit dominiert wurde. Unter den Nukleobasen zeigte sich eine positive Tendenz zugunsten von Thymin und Adenin, während für die Aminosäuren keine so klare physikochemische Präferenz angegeben werden kann [37]. Die Auswahl der Strukturen nahm einen nicht unerheblichen Einfluss auf diese Ergebnisse, da in Doppelhelixstrukturen die innen befindlichen Nukleobasen am wenigsten, die Pentoseringe etwas mehr und die Phosphateinheiten am besten von außen zugänglich sind. Mit der Einführung zusätzlicher einsträngiger Sekundärstrukturelemente verändert sich auch der Expositionsgrad der einzelnen Bestandteile der Nukleotide zugunsten der Nukleobasen. Die fehlende Absättigung der funktionellen Gruppen in den ungebundenen Bereichen derartiger Sekundärstrukturelemente erhöht ferner die Wahrscheinlichkeit einer spezifischen Bindung. Eine unabhängige Studie an 77 Protein-RNA-Komplexen zeigt trotz höherer Zugänglichkeit der Nukleobasen eine ähnliche Verteilung spezifischer und unspezifischer Wasserstoffbrücken. Die detailliertere Aufschlüsselung der Kontakte erlaubt hier jedoch den Rückschluss auf den hohen Grad spezifischer Bindung auf Seiten der Aminosäuren [43].

Bei den aromatischen, hydrophoben und elektrostatischen Wechselwirkungen zeigt sich ebenfalls ein großer unspezifischer Anteil. Da sich die Nukleobasen in Bezug auf diese drei Interaktionsformen kaum unterscheiden, ist die Bindung auf ihrer Seite stets unspezifisch. Anders ist es bei den Seitenketten der Aminosäuren, die sich hier gruppenweise deutlich voneinander unterscheiden. Da die fünf Nukleobasen eine hydrophobe Neigung aufweisen, wird ihre Zusammenlagerung mit hydrophoben Aminosäuren im Sinne des hydrophoben Effekts begünstigt. Ihre Heterozyklen sind in der Lage, Stapelwechselwirkungen mit den aromatischen Ringen einiger der Aminosäuren einzugehen. Schließlich besteht durch die negative Ladung der Phosphateinheiten die Möglichkeit zur elektrostatischen Interaktionen mit polaren und besonders mit positiv geladenen Aminosäuren.

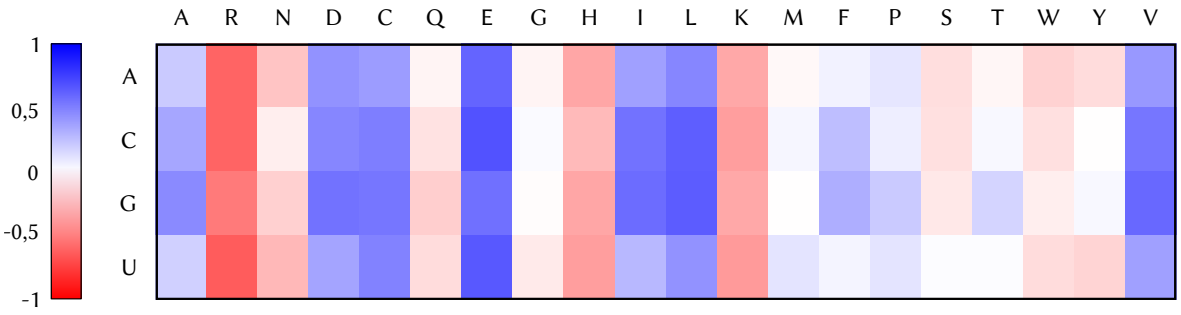


Abb. 2.9: Übersicht der paarweisen Bindungsneigungen zwischen Nukleotiden und Aminosäuren auf einer Skala von bevorzugt (rot) über neutral (weiß) zu benachteiligt (blau), wie sie an einem Beispieldatensatz von 282 Komplexen aus einzelnen Protein- und RNA-Ketten bestimmt wurden. Zwischen den Aminosäuren zeigen sich deutlichere Unterschiede als zwischen den Nukleotiden. Nach [44].

In der graphischen Auftragung der Bindungsneigungen einer dritten Studie, die in Abbildung 2.9 gegeben ist, wird der geringe Gesamteinfluss der spezifischen Interaktionen auf Seiten der Nukleotide und die im Vergleich höhere Spezifität der Aminosäuren deutlich [44]. Der große Anteil unspezifischer Bindungen stabilisiert zwar die Komplexe, führt jedoch in der Abbildung dazu, dass die geringen Unterschiede in der Spezifität der Nukleotide noch schwächer wahrgenommen werden. Zusammenfassend kann festgehalten werden, dass die Bindungsneigungen der einzelnen Aminosäuren und Nukleotide nicht ausreichend sind, um die Bindung eines Protein-Nukleinsäure-Komplexes verwendbar zu charakterisieren. Neben den Sekundärstrukturinformationen sind dafür die geometrischen Details der Tertiärstruktur notwendig.

Unterschiede zwischen DNA und RNA Sowohl DNA als auch RNA können ein- und doppelsträngig vorliegen, wobei im natürlichen Kontext die doppelsträngige DNA und die einzelsträngige RNA dominieren. Ein wichtiger Unterschied besteht ferner in der Form und Zugänglichkeit ihrer helikalen Regionen. Während DNA in der Regel Helices der Form B ausbildet, findet sich aufgrund der veränderten Pentoserestgruppe bei RNA hauptsächlich die Form A. In beiden Formen bilden sich unterschiedlich große Furchen aus, wobei der Größenunterschied in der Form A deutlich ausgeprägter ist. Ihre große Furche ist besonders schmal und tief, sodass Aminosäuren hier kaum eindringen können, um die Nukleobasen zu binden. Die recht breite und flache kleine Furche ist für die Bindungen prinzipiell wesentlich besser geeignet [45]. Die bei DNA fehlende 2′-Hydroxylgruppe der Ribose befindet sich in einer sehr exponierten Position. Sie kann als Akzeptor und Donator für Wasserstoffbrücken dienen und damit die Fähigkeit der unspezifischen Wasserstoffbrückenbildung der RNA-Nukleotide erhöhen.

Die weithin anerkannte Annahme, dass RNA eine prinzipiell höhere Flexibilität aufweist als DNA, muss zugunsten einer komplexeren Definition des Konzepts der Flexibilität verworfen werden. Diese umfasst ohne Anspruch auf Vollständigkeit Subformen der Flexibilität wie Verdrehung, Neigung, Rotation, Verschiebung und Dehnung, die jeweils im lokalen und globalen Bereich betrachtet werden können. So zeigen Helices der beiden Makromolekültypen in verschiedenen Subformen der Flexibilität ein sehr unterschiedliches, relatives Verhalten, sowohl auf lokaler wie auch auf globaler Ebene [46; 47]. Besonders bei Auftreten der einsträngigen Sekundärstrukturelemente ist ein Vergleich der möglichen Flexibilität durch die vielen Freiheitsgrade kaum mehr möglich.

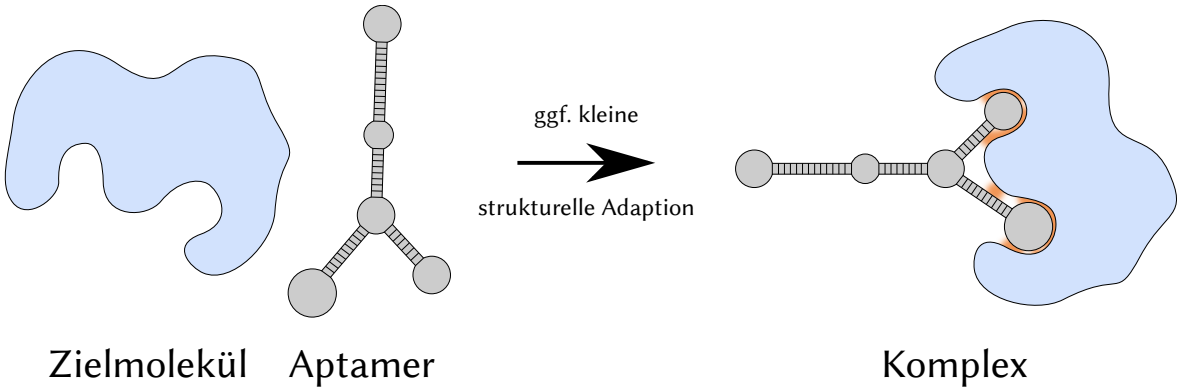


Abb. 2.10: Die Prinzipdarstellung der Aptamer-Target-Bindung zeigt kleine konformationelle Änderungen an der Bindungsfläche des Proteins (blau) und im Verzweigungswinkel des Aptamers (grau) beim Eingehen der Bindung. Auftretende Wechselwirkungen sind orange angedeutet. Nach [48].

2.2 Aptamere und deren Gewinnung

Die soeben eingeführten Grundlagen zu Protein-Nukleinsäure-Komplexen werden im folgenden um eine praktische Anwendung ergänzt. Die im inhaltlichen Schwerpunkt dieser Arbeit liegenden Aptamere sind eine künstlich erzeugte Gruppe funktioneller Nukleinsäuren, deren Interaktion mit Proteinen großes Potenzial in Biologie und Medizin aufweist. Die Vorstellung der Aptamere umfasst dieses Potential ebenso wie die Abgrenzung von ähnlichen Agenten. Schließlich wird das Herstellungsverfahren der Aptamere in seinen Grundzügen vorgestellt, um für die Analyse ein essentielles Verständnis über die methodischen Hintergründe zu fördern und die aufbauenden Optimierungskonzepte in einen sinngerechten Kontext einzubetten. Modifikationen des Grundverfahrens ergänzen schließlich die Betrachtung.

2.2.1 Aptamere als universelle Binder

Das Wort Aptamer setzt sich aus dem lateinischen Bestandteil *aptus* (passend, geeignet, angebunden) und dem griechischen *μερος* (Teil, Ort) zusammen. So wird bereits im Namen der Aptamere ihre prinzipielle Funktion als passende und anbindende Moleküle beschrieben. Die offensichtliche Ähnlichkeit mit Antikörpern beschränkt sich auf die funktionellen Aspekte, denn bei den Aptameren handelt es sich genauer um kurze, einzelsträngige Oligomere aus Nukleotiden, also sowohl RNA als auch *Single Stranded DNA* (ssDNA). Eine schematische Darstellung ist in Abbildung 2.10 gegeben. In ihrer charakteristischen dreidimensionalen Faltung sind sie in der Lage, andere Moleküle zu binden und dadurch zu detektieren oder in ihrer Funktion zu beeinflussen. Über eine geeignete chemische Kopplung können Aptamere auch zum gezielten Transport von Stoffen eingesetzt werden. Das Spektrum molekularer Ziele reicht dabei von einfachen Molekülen bis hin zu ganzen Organismen. In einzelnen Fällen wurden innerhalb des Aptamers Bindungstaschen beobachtet, die ein kleines Zielmolekül durch eine bindungsinduzierte Konformationsänderung einschlossen. Verantwortlich für die Bindung mit den zahlreichen Zielmolekülen sind neben der physikalischen Passform hauptsächlich die Stapelung aromatischer Ringe, elektrostatische und van-der-Waals-Wechselwirkungen sowie Wasserstoffbrückenbindungen [48; 49]

Bindepartner von Aptameren Ihre zahlreichen Bindemodi und die Vielseitigkeit ihrer Tertiärstrukturen erlauben die spezifische und hochaffine Bindung von Aptameren an eine Vielzahl sehr unterschiedlicher Zielmoleküle. Dies wird am Beispiel bereits durchgeführter Aptamerselektionen deutlich, die exemplarisch und ohne Anspruch auf Vollständigkeit in der folgenden Auflistung umrissen werden. Als Vertreter der anorganischen Stoffe wurden aufgrund der eigenen Ladung der Aptamere ausschließlich positiv geladene Metallionen genutzt [50–53]. Von deutlich höherem Interesse sind jedoch die organischen Zielverbindungen. Im Bereich der eher kleinen organischen Moleküle wurden Aptamere unter anderem für natürlich auftretende Toxine [54–56], Hormone und andere hormonell aktive Stoffe [57–60], einige häufig missbrauchte psychotrope Substanzen [61; 62], Farbstoffe [63–65] und weitere nicht in diese Klassifikation einzuordnende Vertreter [66–69] selektiert. Auch unter den großen organischen Molekülen und deren Grundbausteinen finden sich zahlreiche Ziele für die Selektion von Aptameren. Hier sind sowohl Nukleotide [70; 71] und Nukleinsäuren [72; 73], Kohlehydrate [74–76], als auch Aminosäuren [77–79], Peptide und Proteine [80–84] als wichtige Vertreter der kettenförmigen Makromoleküle zu nennen. Schließlich wurden auch Kofaktoren [85; 86] und Antibiotika [87–89] in den Fokus der Aptamerentwicklung genommen. Wenn auch durch die erschwerten experimentellen Bedingungen zur Zeit noch die Ausnahme, so konnten auch bereits Aptamere für sehr komplexe und große Ziele selektiert werden. Im Gegensatz zu zahlreichen Ansätzen, die auf der Erkennung einzelner spezifischer Bestandteile des Gesamtzieles basieren, existieren auch einzelne, die die gesamte Makrostruktur zur Selektion der Aptamere einsetzten. Neben Viren [90; 91], Sporen [92; 93] und Zelloberflächen [94; 95] sind hier besonders die Ergebnisse mit lebenden Einzellern als Zielstruktur [96; 97] hervorzuheben.

Aptamere eignen sich jedoch als Binder nicht für alle molekularen Strukturen gleich gut. Die intermolekularen Wechselwirkungen im Aptamer-Zielmolekül-Komplex sind abhängig von den physikochemischen Eigenschaften der Komplexpartner und deren Miteinander. Durch die negativen Ladungen der Phosphatgruppen des Nukleinsäurerückgrats weisen Aptamere etwa ein wesentlich besseres Bindeverhalten zu positiv als zu negativ geladenen Zielstrukturen auf. Ferner ermöglichen Donatoren und Akzeptoren für Wasserstoffbrücken sowie aromatische Ringstrukturen entsprechende Interaktionen mit dem Aptamer. Moleküle mit überwiegend hydrophobem Charakter oder entsprechenden Gruppen an wichtigen Bindepunkten behindern im Gegensatz dazu die Ausbildung einer Aptamerbindung [48], was jedoch nicht als totales Ausschlusskriterium zu interpretieren ist [98]. Zur Durchführung der Selektion muss schließlich sichergestellt werden können, dass die Zielmoleküle oder -strukturen in hinreichender Menge und Reinheit zur Verfügung gestellt werden können.

Anwendungsbereiche für Aptamere In Verbindung mit ihrer Funktion als universelle Binder ergeben sich für Aptamere eine Reihe von praktischen Anwendungsfeldern. Durch das Andocken an ein Zielmolekül kann dieses für einen Folgeprozess markiert [99–101] oder in Form eines fixierten Aptamers innerhalb eines Biosensors detektiert werden [102–104]. Auch das Einfangen spezifischer Zielmoleküle [105; 106] im Rahmen eines Filterprozesses ist möglich. Die molekulare Bindung besitzt jedoch über diese Fälle hinaus auch das Potential, eine biologische Wirkung auszuüben. Diese lässt sich in zwei grundlegende Kategorien einteilen, die direkte Wirkung des Aptamers und die Unterstützung der Wirkung Dritter. Ist die Zielstruktur biologisch aktiv, so kann das Aptamer durch die Belegung des aktiven Zentrums oder durch Induktion einer strukturellen Anpassung die natürliche Funktion der Zielstruktur graduell inhibieren oder gänzlich unterbinden [107–109]. Weist das Aptamer hingegen ein ähnliches Interaktionsmus-

ter wie der natürliche Interaktionspartner der Zielstruktur auf, so besteht die Möglichkeit eine Aktion der Zielstruktur durch das Aptamer zu triggern. So konnte beispielsweise mit einem postselektiv modifizierten Aptamer ein Rezeptor erfolgreich ausgelöst werden [110; 111]. Ein anderer Ansatz wird beim sogenannten *Drug Delivery* verfolgt. Hier wird eine Nutzlast an das Aptamer angebracht, die anschließend mit dem Aptamer gemeinsam an die spezifische Zielposition gelangt. Diese Nutzlast kann ein Wirkstoff sein, der am Zielort direkt seine Wirkung entfaltet, oder ein Hilfsstoff, der dort einen von außen gesteuerten Prozess unterstützt [112; 113]. Die charakteristischen Eigenschaften der Aptamere erschließen dabei neue Möglichkeiten der örtlichen Zugänglichkeit, die von großem Interesse sein können. Hier sei exemplarisch erwähnt, dass unter Laborbedingungen an Mäusen die Blut-Hirn-Schranke mit Aptameren durchbrochen werden konnte [114]. Auch wenn dieser Grad der Zugänglichkeit noch nicht mit der Selektion auf eine konkrete Zielstruktur hin kombiniert wurde, zeigt diese Studie ein großes Potential der Aptamere.

Gemeinsam betrachtet mit der Fülle möglicher Zielmoleküle ergeben sich für Aptamere daher interessante Perspektiven in Medizin und Industrie. Mit Industriezweigen wie der Nahrungsmittelkontrolle, der Biotechnologie und der Schadstofffilterung bietet sich für die Aptamere ein breites Anwendungsfeld mit hoher Relevanz für das menschliche Leben. In der Medizin bleiben den Aptameren weder die Richtungen Diagnose noch Therapie verschlossen, wie sich an aktuellen Forschungsvorhaben offenbart. Mit nur einem zugelassenen und elf weiteren in der klinische Prüfung befindlichen Produkten zeigt sich jedoch auch, dass der praktische Nutzung der Aptamertechnologie in der Medizin einige Hürden im Weg stehen. Neben dem komplexen Prüfungsprozess und der hohen Verantwortung spielt die Wirkungsumgebung hier eine große Rolle. Die tatsächlichen Umgebungsbedingungen können dem *in vitro* selektierten Aptamer *in vivo* große Probleme bereiten. So können ungewollte Bindungen mit der komplexen Matrix strukturelle und damit auch funktionelle Änderungen induzieren oder Bindebereiche blockieren. Zudem können körpereigene Degradationsmechanismen für Nukleinsäuren die Aptamere angreifen. Da für den medizinischen Einsatz spezifische Bindestellen auf der Zielstruktur entscheidend sind, welche in einem Selektionslauf nicht in jedem Fall beeinflusst werden können, muss für die Gewinnung von Aptameren hier außerdem mit einem höheren Arbeitsaufwand gerechnet werden [115; 116].

Vergleich von Aptameren und Antikörpern Aptamere und Antikörper unterscheiden sich in ihrem Aufbau aus Nukleotiden und Aminosäuren, haben jedoch einen großen Teil ihrer Anwendungsfelder gemeinsam. Entsprechend kommt es zu einer Konkurrenz der beiden Molekültypen. Ein bedeutender Vorteil der Aptamere ist ihre geringere Größe und die daraus resultierende höhere Zugänglichkeit zu biologischen Kompartimenten. Sowohl bezogen auf die Affinität als auch auf die Spezifität konnten mit Aptameren ähnliche und bessere Ergebnisse erzielt werden als mit Antikörpern. Dies mindert ungewollte Kreuzreaktionen und damit im praktischen Umfeld die zu erwartenden Nebenwirkungen und die benötigte Dosis. Auch in Bezug auf Immunogenität und Toxizität zeigen die Aptamere positivere Eigenschaften als Antikörper. Nach der initialen Selektion und Validierung eines Aptamers erfolgt seine Erzeugung auch in großen Mengen durch die chemische Synthetisierung wesentlich schneller und kostengünstiger als die eines vergleichbaren Antikörpers. Die Variationen der Chargen sind minimal und sowohl bakterielle als auch virale Kontaminationen stellen kaum mehr ein Problem dar, da keine lebenden Organismen zur Produktion erforderlich sind. Die Lagerung von Aptameren ist im Vergleich zu Antikörpern auch über längere Zeiträume hinweg unkompliziert möglich. Sowohl während

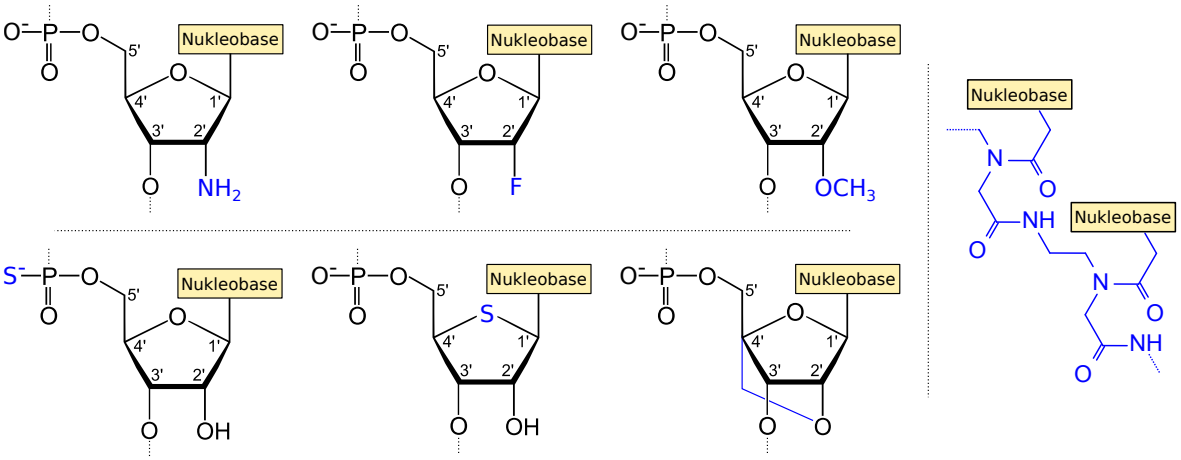


Abb. 2.11: Darstellung einer Auswahl möglicher Modifikationen des Aptamers im Bereich des Zucker-Phosphat-Rückgrats. Die im Vergleich zu der in Abbildung 2.6 vorgestellten Grundkonfiguration vorgenommenen chemischen Modifikationen sind blau gekennzeichnet. Die tabellarische Anordnung kleiner Modifikationen im linken und mittleren Bildbereich zeigt in der oberen Reihe die Modifikationen an der 2′-Position des Pentoserings und in der unteren Reihe Schwefelmodifikationen an Phosphat und Pentosering sowie die kovalente Bindung der LNA. Im rechten Bildbereich ist die Grundstruktur der PNA dargestellt, bei der die Nukleobasen an einem Peptidrückgrat befestigt sind.

der Lagerung als auch in der Verarbeitung stellen temporäre Denaturierungseffekte kein praktisches Problem dar, da über einfache thermische Verfahren eine Renaturierung der Aptamere erreicht werden kann. Durch die körpereigenen Abbauprozesse sind zumindest unmodifizierte Aptamere wesentlich stärker betroffen als Antikörper und modifizierte Aptamere. Dies führt zu einem schnelleren Abbau und damit zu einer verkürzten Wirkdauer [48; 117].

Chemische Modifikationen an Aptameren Zur Verbesserung der physikochemischen Eigenschaften, besonders aber ihrer Beständigkeit gegen die natürlichen Degradationsmechanismen steht ein Repertoire chemischer Modifikationen für Aptamere zur Verfügung. Es ist zu beachten, dass einige der Modifikationen erst nach Abschluss der Selektion durchgeführt werden können, da die modifizierten Aptamere von den im Selektionsverfahren eingesetzten Polymerasen ansonsten nicht als Substrat erkannt werden. Als Ausgangspunkt für die Modifikationen gilt dabei stets der in Abbildung 2.6 gezeigte, grundlegende Aufbau der Nukleotide. Er erlaubt die Modifikation prinzipiell an zwei Stellen. Mit dem Einsatz modifizierter Nukleobasen wird dabei sehr großer Einfluss auf die sequenziell bestimmten Bindeeigenschaften entlang des Hauptstranges ausgeübt. Da dies zu Weit von der Zielstellung dieser Arbeit abweicht, soll es nur am Rande erwähnt bleiben [115]. Im Bereich des Nukleinsäurerückgrats können Modifikationen vorgenommen werden, die die physikochemischen Eigenschaften der Nukleobasen nicht beeinflussen und damit mit großer Wahrscheinlichkeit die Bindefähigkeit der modifizierten Aptamere nur geringfügig beeinträchtigen. Diese Modifikationen werden im folgenden kurz vorgestellt und sind in Abbildung 2.11 als Übersicht gegenübergestellt.

Um die Degradationsresistenz zu erhöhen, greifen zahlreiche Ansätze an der 2′-Position der Pentose an. Hier hat sich neben weiteren die Substitution der vorhandenen Hydroxygruppe durch eine Aminogruppe, Fluor oder auch eine Sauerstoff-Methylgruppe (OCH_3) als wirkungsvoll erwiesen. Innerhalb der Pentose führt die kovalente Verbindung des 2′-Sauerstoff mit dem 4′-Kohlenstoff über eine Methylenbrücke zu einer erhöhten thermischen Stabilität und geringeren Anfälligkeit des Aptamers für falsche Basenpaarungen. Derartig modifizierte Aptamere wer-

den auch *Locked Nucleic Acid* (LNA) genannt und sind zusätzlich gegen die Degradation durch Nukleasen resistent. Weiterhin wurde mit der Substitution von Sauerstoff durch Schwefel sowohl im Pentosering als auch in der Phosphateinheit des Nukleinsäurerückgrats experimentiert. Besonders die Ersetzung des nicht-bindenden Sauerstoffatoms in der Phosphodiesterbindung zeigte hier Erfolg. Die entstehende Phosphorothioatbindung erhöht neben der Beständigkeit des Aptamers gegen Zersetzung durch Nukleasen auch seine Fähigkeit, die Doppellipidschicht von Biomembranen zu durchdringen [115; 118].

Zahlreiche Exonukleasen zeigen ein gerichtetes Verhalten, initiieren ihre degradierende Aktivität also spezifisch an einem Ende des Nukleinsäurestranges. In diesem Zusammenhang konnte gezeigt werden, dass der Abbau von Aptameren unterbunden werden kann, indem das für die Exonuklease relevante Ende des Nukleinsäurestranges bedeckt wird. Praktisch wird zu diesem Zweck beispielsweise ein invertiertes Nukleotid eingesetzt, sodass das Aptamer zwei gleiche Enden ausweist. Sind sowohl 3′-5′- als auch 5′-3′-Exonukleasen vorhanden, so können auch andere Bedeckungen gewählt werden. Strukturell weitreichendere Modifikationen umfassen den Austausch des gesamten Nukleinsäurerückgrats durch andere molekulare Gerüstsysteme. Hier seien besonders die *Peptide Nucleic Acid* (PNA) erwähnt, die durch den Einsatz des Peptidrückgrates weder von Nukleasen noch von Proteasen angegriffen werden. Die veränderten Ladungsverhältnisse des Rückgrats wirken sich jedoch auf das elektrostatische Bindeverhalten aus [115; 119]. Eine völlig andere Herangehensweise ist die der Spiegelmere. Sie setzen ausschließlich die nicht-natürlich vorkommenden L-Enantiomere der Nukleotide ein, die von den natürlichen Degradationsmechanismen nicht betroffen sind. Da auch die im konventionellen Selektionsverfahren eingesetzten Enzyme nicht auf diese Enantiomere ansprechen, erfolgt ihre Generierung über einen Umweg mithilfe von gespiegelten Zielmolekülen und konventionellen Aptameren [120].

2.2.2 Das Grundverfahren der Aptamerselektion

Als Grundverfahren für die Aptamerselektion gilt das iterative *Systematic Evolution of Ligands by Exponential enrichment* (SELEX), welches erstmals 1990 durch Craig Tuerk und Larry Gold beschrieben wurde. Aus einer zu prüfenden Menge von Oligonukleotiden ermöglicht es die Auswahl derer Kandidaten, die in der Lage sind, ein bestimmtes Zielmolekül mit der höchsten Affinität zu binden [121]. Für die Aptamerselektion mit dem Verfahren SELEX ergibt sich der in Abbildung 2.12 skizzierte, prinzipielle Ablauf. Der Ausgangspunkt eines typischen SELEX-Prozesses ist eine chemisch synthetisierte Bibliothek von großteils zufällig zusammengesetzten Oligonukleotiden. Trotz ihrer enormen Größe von bis zu 10^{16} Oligonukleotiden deckt die Bibliothek nur einen geringen Anteil des Sequenz- und damit des Faltungsraumes der Oligonukleotide ab. In mehreren aufeinanderfolgenden Runden werden nun die Oligonukleotide der aktuellen Bibliothek mit den jeweiligen Zielmolekülen inkubiert, sodass diejenigen Vertreter mit relativ hoher Affinität eine Bindung mit dem Zielmolekül eingehen können. Häufig wird dabei der künstlich erzeugte Selektionsdruck über die Runden hinweg sukzessive erhöht. In einer Separationsphase werden alle nicht-gebundenen Oligonukleotide abgetrennt und entfernt. Die mit unterschiedlicher Affinität gebundenen Kandidaten werden schließlich von den Zielmolekülen eluiert und durch Amplifikation sowie Reinigung für die nächste Runde des Experiments aufgearbeitet. Die immobilisierten Zielmoleküle können dabei in der Regel wiederverwendet werden. Das Fortschreiten der Selektion ist sowohl von der Bindefähigkeit des Zielmoleküls, der Zusammensetzung der Ausgangsbibliothek als auch von den eingestellten Umgebungsbedingungen stark abhängig. Die an den unterschiedlichen Stadien des Experiments eingesetzten

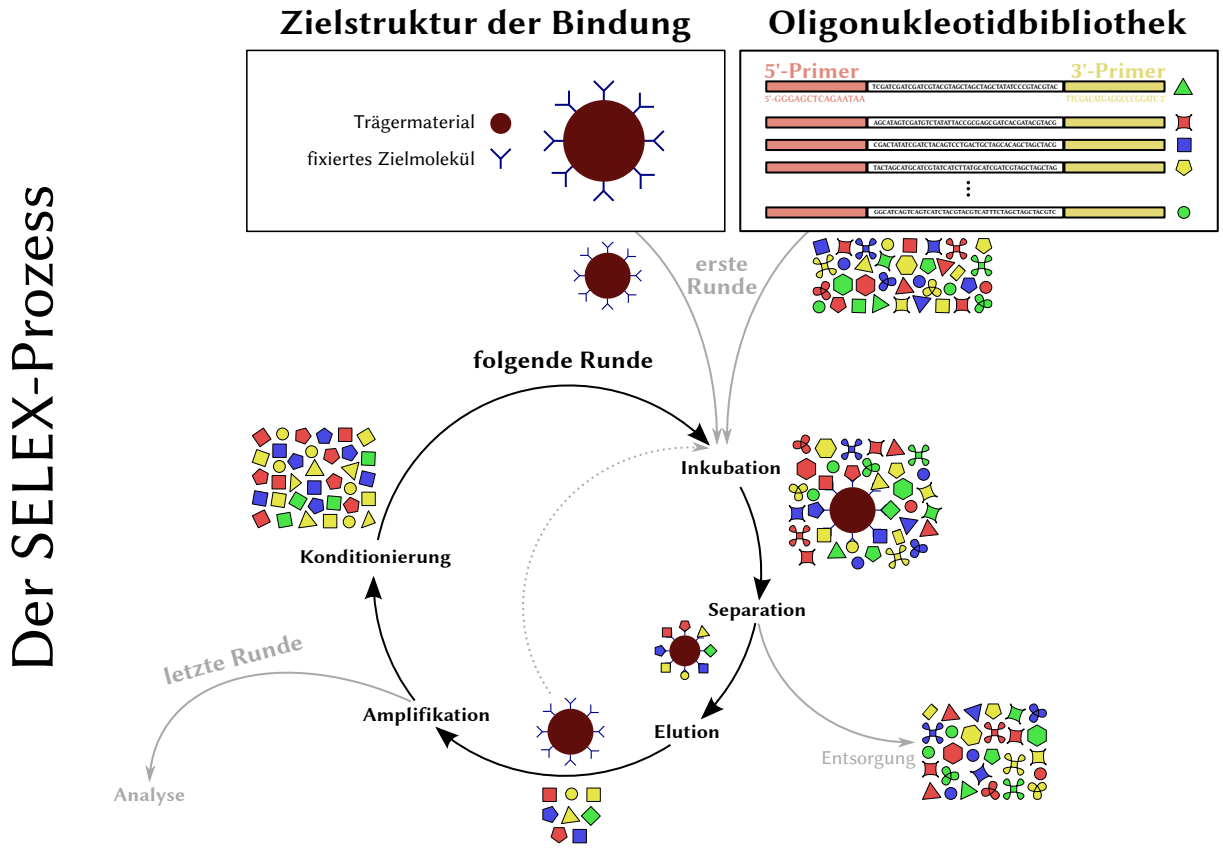


Abb. 2.12: Übersichtsdarstellung des in mehreren Runden ablaufenden Selektionsverfahrens SELEX für Aptamere. Die im oberen Bereich der Abbildung angeordnete Zielstruktur und Oligonukleotidbibliothek werden in die initiale Runde des SELEX-Prozesses eingebracht. Jede Runde des Experiments besteht aus einer Inkubationsphase der Oligonukleotide mit den fixierten Zielmolekülen, der Separation und Entfernung ungebundener Vertreter, der Elution (Ablösung) und Amplifikation der gebundenen Kandidaten sowie deren Konditionierung. Die Zielstrukturen können dabei oft wiederverwendet werden. Nach Abschluss der letzten Runde werden die verbliebenen Kandidaten in der Regel anschließenden Analysen unterzogen. Alle nicht zum Kernzyklus gehörenden Pfeile sind schwächer gezeichnet. Nach [48].

Methoden und deren Konfiguration übt hingegen einen untergeordneten Einfluss aus. Trotz dieser vielen Faktoren hat sich in der Praxis eine Anzahl von 6 bis 20 Runden als ausreichend herausgestellt. Nach dem Ende des SELEX-Experimentes und einer finalen Vervielfältigung werden die Aptamerkandidaten zur weiteren Analyse bereitgestellt [48].

Erzeugung einer initialen Oligonukleotidbibliothek Der Ausgangspunkt eines SELEX-Prozesses ist die chemisch synthetisierte Oligonukleotidbibliothek, deren Sequenzen sich in der Regel dreigliedrig zusammensetzen. Der zentrale, randomisierte *Insert*-Bereich mit einer Länge von typischerweise 20 nt bis 80 nt wird dabei jeweils beidseitig von etwa 18 nt bis 21 nt langen *Sense*- und *Antisense*-Primersequenzen umschlossen, die für die spätere Amplifikationsphase notwendig sind. Sequenzen außerhalb dieses Größenrahmens werden kaum eingesetzt, da sie dann entweder über eine zu geringe strukturelle Vielfalt verfügen oder durch sinkende Abdeckung des Sequenz- und Strukturraumes kaum mehr Gewinn aus der zunehmenden Größe gezogen werden kann. Eine durchschnittliche Oligonukleotidbibliothek enthält dabei etwa 10^{13} bis 10^{16} Einzelsequenzen, wobei die experimentellen Rahmenbedingungen wenig Raum zur Verwendung einer größeren Bibliothek lassen. Diese Dimensionierung entspricht im Vergleich etwa der des kompletten Sequenzraumes von Aptameren einer Länge von 22 nt bis 27 nt. Über-

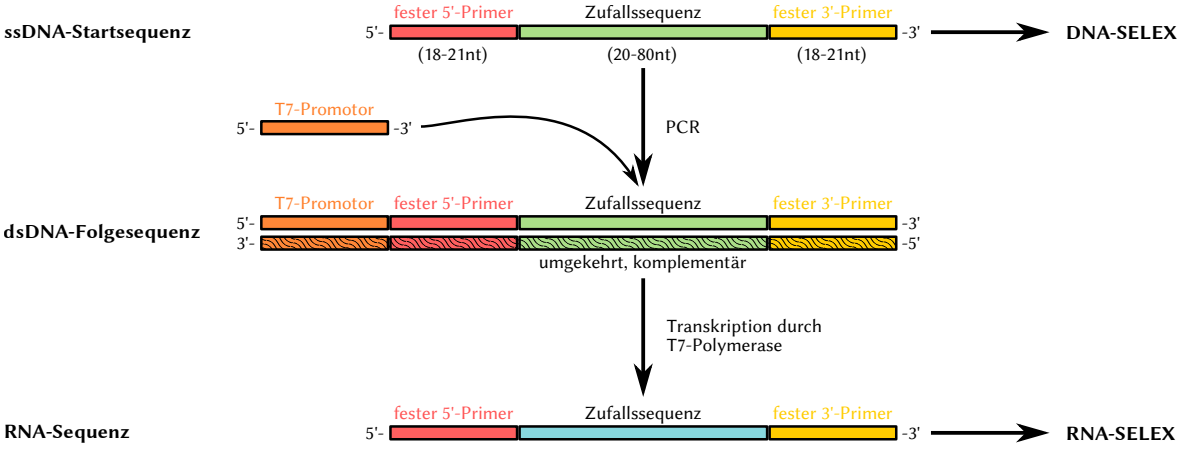


Abb. 2.13: Aufbau und Erzeugung von Oligonukleotidbibliotheken mit unterschiedlicher Verfahrensweise für DNA- und RNA-Bibliotheken. Die Zufallssequenz (grün für DNA und blau für RNA) wird von zwei Primern umschlossen (rot und gelb). Für die Erzeugung der RNA-Bibliothek ist eine PCR und anschließende Transkription notwendig. Nach [48].

steigt die Aptamerlänge diesen Bereich, so kann der mögliche Sequenz- und Strukturraum trotz der großen Dimension der Bibliothek nur zu einem sehr kleinen Bruchteil abgedeckt werden. Hinzu kommt, dass auch auf Basis der zufälligen Synthetisierung sowohl Doppelungen als auch Lücken in der Abdeckung des Sequenz- und Strukturraums entstehen. Entsprechend muss bei zahlreichen Aptamerselektionen davon ausgegangen werden, dass die genutzte Bibliothek nicht die optimalen Aptamersequenzen enthält, sondern nur Vertreter mit suboptimaler Affinität zum Zielmolekül. Das Standardverfahren sieht einen vollständig randomisierten inneren Bereich für die Aptamere vor, wobei die Nukleotidkomposition in der Synthetisierung angepasst werden kann. Neueste Erkenntnisse zeigen, dass kommerziell erhältliche Oligonukleotidbibliotheken häufig mit einem, wenn auch geringen, Bias in der Nukleotidkomposition ausgeliefert werden. Diese Bias wirkt sich nicht nur auf die Nukleotidkomposition der finalen Aptamere aus, sondern auch auf die Gesamteffizienz des Selektionsverfahrens [122].

Die praktischen Schritte zur Erzeugung einer Bibliothek unterscheiden sich je nach eingesetztem Nukleinsäuretyp, wie in Abbildung 2.13 verdeutlicht wird. Während eine DNA-Bibliothek direkt nach der Synthetisierung verwendet werden kann, ist für die Erzeugung einer RNA-Bibliothek ein Umweg über die zugehörige ssDNA notwendig. Diese nur temporäre ssDNA wird dazu mit einem speziellen *Sense*-Primer ausgestattet, der die Promotorsequenz der T7-RNA-Polymerase als Erweiterung an seinem $5'$-Ende trägt. Durch eine *Polymerase Chain Reaction* (PCR) werden die Sequenzen dann in *Double Stranded DNA* (dsDNA) überführt und schließlich mithilfe der T7-RNA-Polymerase *in vitro* zu RNA transkribiert [48].

Selektionsphase Der Begriff Selektionsphase fasst die drei einzelnen Prozessschritte der Inkubation, Separation und Elution zusammen, durch die aus der zugrunde liegenden Anfangsbibliothek der jeweiligen SELEX-Runde Aptamerkandidaten mit großer Affinität zum Zielmolekül freigestellt werden. Die Selektion beginnt mit der Inkubationsphase, in der die Zielmoleküle mit der Oligonukleotidbibliothek zur Bindung gebracht werden. Die Umgebungsbedingungen sind dabei stark von den gewählten Bibliotheken und Zielmolekülen abhängig und sollen daher an dieser Stelle nicht weiter spezifiziert werden. Über die Begrenzung der Inkubationsdauer und der Menge an eingesetzten Zielmolekülen wird ein künstlicher Selektionsdruck auf die

Aptamerkandidaten geschaffen. Dieser wird in der Regel zu Beginn niedrig angesetzt, um die Bandbreite möglicher Binder groß zu halten. In den folgenden Runden wird der Selektionsdruck schließlich erhöht, um Kandidaten mit hoher Affinität stärker zu fördern [48; 123]

Nach der Inkubation erfolgt die Separation von gebundenen und ungebundenen Oligonukleotiden. Lassen sich die Zielmoleküle auf einer speziellen Matrix immobilisieren, so reicht hier ein Waschschritt aus, um die Fraktionen optimal aufzutrennen. Ist dies nicht der Fall, werden die folgenden Methoden für die Separation verwendet. Ein Mittel der Wahl ist die Affinitätschromatographie [124], die jedoch zur effizienten Beladung der Säulen erhebliche Mengen der Zielmoleküle erfordert. Im Falle geringer Mengenverfügbarkeit muss daher auf Verfahren wie *Magnetic Beads* [125] ausgewichen werden, welche über die Fixierung mit magnetischen Partikeln einfach handhabbar sind und geringere Mengen an Zielmolekülen voraussetzen. Ist keinerlei Immobilisierung der Zielmoleküle möglich oder aufgrund damit verbundener Einflüsse auf deren Faltung nicht erwünscht, so können die Oligonukleotide nur mit geringerer Präzision separiert werden. Hier finden neben der gewichtsbasierten Ultrafiltration noch andere Verfahren wie beispielsweise die Kapillarelektrophorese [54], die Durchflusszytometrie [126], die Oberflächenplasmonenresonanzspektroskopie [127] und die Zentrifugierung [128] Anwendung. Handelt es sich bei den Zielmolekülen um Proteine, so bietet sich ferner die Möglichkeit der Filterung durch eine Nitrozellulosemembran [129], an der die Proteine haften. Über radioaktive [130] oder fluoreszierende [125] Marker können die gebundenen und ungebundenen Mengenanteile der Oligonukleotide während des Experiments beobachtet werden [48; 123]. Aktuelle Fortschritte in der Hochdurchsatzsequenzierung erlauben nachträglich zudem die Beobachtung der Entwicklung der Bibliothek über die einzelnen Runden hinweg [131].

Für die weitere Verarbeitung müssen nun im Schritt der Elution Zielmoleküle und Oligonukleotide wieder voneinander getrennt werden. Unter Zuhilfenahme von Hitze [125] oder speziellen Substanzen wie Harnstoff, Natriumlaurylsulfat (SDS) oder Ethylendiamintetraessigsäure (EDTA) [132–134] kann dies durch Denaturierung der Oligonukleotide erreicht werden. Die Elution kann jedoch auch über kompetitive Binder [135] ausgelöst werden. Die Wiederverwendung der Zielmoleküle ist nach der Elution im Normalfall möglich [48].

Amplifikation der Ergebnisbibliothek Die Anzahl der Oligonukleotide ist nach der Selektionsphase zu stark reduziert, um direkt in der nächsten Runde eingesetzt zu werden. Aus diesem Grund werden die verbliebenen Kandidaten amplifiziert, wozu für ssDNA-Stränge lediglich die Anwendung einer PCR notwendig ist. Eine RNA-Bibliothek muss vor diesem PCR-Schritt durch eine *Reverse Transcription PCR* (RT-PCR) in ihre komplementären ssDNA-Stränge überführt werden, wozu analog zur Bibliothekserzeugung die Promotorsequenz der T7-RNA-Polymerase erforderlich ist [48]. Das Ergebnis der Amplifikation ist in jedem Fall ein doppelsträngiger DNA-Pool, der im Rahmen der Konditionierung für die nächste Runde aufgearbeitet wird, um wieder einsträngige Sequenzen des richtigen Typs zu erhalten. Zu diesem Zweck werden entweder über eine erneute Transkription mit T7-RNA-Polymerase die entsprechenden RNA-Stränge generiert oder die zwei Stränge der dsDNA voneinander getrennt. Hierzu kann beispielsweise das Streptavidin/Biotin-System in verschiedenen Modi aber auch eine asymmetrische Variante der PCR mit nur einem Primer verwendet werden [136]. Die Fehlerraten der hier beschriebenen enzymatischen Verfahren sind in der Regel hinreichend gering, sodass der Effekt der Mutation an dieser Stelle keine besondere Beachtung erfordert [137].

Abschluss des Experiments Der SELEX-Prozess endet in der Regel nach 6 bis 20 Runden, wenn die Diversität der Bibliothek durch Anreicherung hochaffiner Aptamerkandidaten hinreichend abgenommen hat, da eine weitere Entwicklung der Bibliothek ohne Mutation ausgeschlossen werden kann. Ein wichtiger Schritt für die anschließende Auswertung ist die Bestimmung der genauen Affinitäten k_D der gewonnenen Aptamere zum Zielmolekül. Aufgrund des experimentellen Aufwandes beschränkt sich dies in der Regel nur auf eine Auswahl der finalen Bibliothek. Ein Großteil der chemischen Modifikationen der Aptamere kann ebenfalls erst nach Abschluss der Selektion durchgeführt werden.

In der Regel erfolgt nach Abschluss der Selektion die Sequenzierung einer kleinen Anzahl von hoch angereicherten Aptamerkandidaten der finalen Bibliothek. Dem kann sich eine experimentelle Verifikation der Bindung zwischen Aptamer und Zielmolekül sowie die Suche nach Gemeinsamkeiten mithilfe eines Multisequenzalignments anschließen [48]. Mit Technologien der Hochdurchsatzsequenzierung wie *Next Generation Sequencing* (NGS) kann darüber hinaus nahezu die gesamte Bibliothek sequenziert und einer weiterführenden bioinformatischen Analyse unterzogen werden [131]. Die Datensätze einer solchen Sequenzierung erfordern mit ihrer Größe im Bereich von 10^4 bis 10^6 Sequenzen komplexere Analyseverfahren als das Multisequenzalignment. Zum Erlangen eines exakten Verständnisses der Bindung zwischen Aptamer und Zielmolekül sind jedoch dreidimensionale Strukturen der beiden Bindepartner und des Komplexes notwendig, aus denen die notwendigen atomaren Kontakte und Bindecharakteristika hinreichend genau hervorgehen. Die Bestimmung dieser Strukturen über Verfahren wie Röntgenkristallographie und Kernspinresonanzspektroskopie gestaltet sich jedoch aufgrund der experimentellen Anforderungen schwierig [138]. Ohne derartige Strukturen geht die Suche nach einer minimalen, stabilen Aptamervariante meist den Weg über im Alignment konservierte Sequenzmotive und experimentelle Analysen auf deren Basis gekürzter Aptamere [48].

2.2.3 Methodische Modifikationen des Selektionsverfahrens

Basierend auf dem eben vorgestellten Grundverfahren von SELEX wurden zahlreiche spezifische Adaptionen und Erweiterungen entwickelt. Während das ursprüngliche Ablaufprinzip erhalten bleibt, optimieren die einzelnen Ansätze die Spezifität und Universalität der gefundenen Aptamere oder die Effizienz des Verfahrens an sich. Eine exemplarische Auswahl ohne Anspruch auf Vollständigkeit soll hier kurz vorgestellt werden.

Erhöhung der Spezifität Die mit SELEX gewonnenen Aptamere weisen in der Regel eine hohe Affinität auf. Durch die ständige Präsenz experimentell notwendiger Hintergrund- und Trägermaterialien kann es jedoch auch zur Anreicherung von Aptameren mit ungewollter Hintergrundbindung kommen. Da der experimentelle Ablauf der Selektion zwar die Affinität der Aptamere, nicht jedoch ihre Spezifität zum Zielmolekül optimiert, besteht die Möglichkeit, dass die gefundenen Aptamere auch Analoga ihres Zielmoleküls binden. Ist dies nicht erwünscht, so kann der Prozessablauf des Grundverfahrens zugunsten einer Selektion spezifischer Aptamere erweitert werden. Die Form des *Negative SELEX* [139] führt zwischen den normalen Runden zusätzliche Negativselektionen ein, in denen die Oligonukleotidbibliothek mit den Hintergrund- und Trägermaterialien in Abwesenheit von Zielmolekülen inkubiert wird. Nach der Separation werden diejenigen Vertreter, die nicht an diese Matrix gebunden haben, der nächsten Runde zugeführt. Die Hintergrundbinder werden damit entfernt. Aufbauend auf *Negative SELEX* werden in den Negativselektionsrunden der Form *Counter SELEX* [67] Moleküle eingebracht, die dem Zielmolekül ähnlich sind, von den Aptameren jedoch nicht gebunden werden sollen. Durch das

anschließende Entfernen derjenigen Aptamerkandidaten, die an diese Gegenziele banden, wird eine hohe Spezifität für das eigentliche Zielmolekül erreicht. Diese Methode ist besonders geeignet für kleine Zielmoleküle, deren Analoga bekannt sind. Im Falle komplexer Zielstrukturen kann sich die Spezifität auf bestimmte Bereiche der Oberfläche beziehen. Zur Negativselektion werden daher Vertreter der Zielstruktur erforderlich, die den gewünschten Teilbereich der Oberfläche nicht enthalten. Der ansonsten dem *Counter SELEX* gleichende Ansatz wird in Zusammenhang mit komplexen Zielstrukturen *Subtractive SELEX* [140] genannt [141].

Erhöhen der Universalität Es besteht ferner das Bestreben, die Universalität des SELEX-Verfahrens zu erhöhen, indem weitere komplexe oder in physikochemischer Sicht problematische Zielmoleküle für die Aptamerselektion erschlossen werden [98; 142; 143]. In diesem Abschnitt sollen jedoch Verfahren angerissen werden, die die Universalität der gewonnenen Aptamere fokussieren. Es existieren Anwendungsfälle, in denen Aptamere nicht nur ein einzelnes Zielmolekül erkennen müssen, sondern eine Reihe von analogen Molekülen. Auch wenn dies bereits als Nebeneffekt im Grundverfahren stattfinden kann, ist ein gezieltes Vorgehen notwendig, um diesen Effekt sicherzustellen. Das Verfahren *Toggle SELEX* [132; 144] baut dabei auf den alternierenden Einsatz mehrerer Zielstrukturen. Auf diese Weise können nicht nur universellere Aptamere geschaffen, sondern auch gemeinsame Epitope der Zielstrukturen fokussiert werden. Vereinzelt wurden zudem Versuche unternommen, den Selektionsvorgang in gewisser Analogie zum Grundverfahren in lebenden Zellen durchzuführen [145]. Da hier die Umgebungsbedingungen des späteren Einsatzes bereits während der Selektion vorherrschen, ist mit einer hohen praktischen Anwendbarkeit der Aptamere zu rechnen. Die Umsetzung in lebenden Zellen ist jedoch mit großen Schwierigkeiten verbunden. Schließlich widmeten sich Ansätze wie *Chimeric* und *Multi Stage SELEX* [146; 147] der Untersuchung von Aptamerkombinationen. Nach der Selektion verschiedener Aptamere für eine spezielle Zielstruktur erfolgt dabei die Fusionierung der erhaltenen Aptamere. Über weitere Selektionsrunden werden schließlich die erhaltenen Aptamerkombinationen evaluiert [148].

Veränderung der Startbibliotheken Ein weiterer wichtiger Ansatzpunkt für Optimierungen ist die Oligonukleotidbibliothek des Verfahrens. Durch Einsatz von modifizierten Nukleotiden [149], die mit den restlichen Prozessschritten des Selektionsverfahrens kompatibel sind, kann eine Bibliothek erzeugt werden, deren Aptamere eine höhere Degradationsstabilität oder veränderte physikochemische Eigenschaften aufweisen. Es wurden jedoch auch einzelne Versuche [150–152] unternommen, durch die teilweise Vorgabe eines abstrakten Motivs innerhalb der randomisierten Bibliothek bestimmte Sequenz- und Strukturelemente in die Aptamere fest zu integrieren, deren Nützlichkeit bekannt oder vermutet war. Derart strukturierte Bibliotheken führten zu guten Ergebnissen [48; 137].

Im *Blended SELEX* [153] werden spezifische Moleküle kovalent an die Oligonukleotide der Bibliothek gebunden, die selbst keine Nukleotide sind. Diese können durch zusätzliche funktionelle Gruppen helfen, das Aptamer gezielt an ein bestimmtes Epitop am Target oder überhaupt mit höherer Affinität anzulagern. Auf ähnliche Weise wurde auch der Einsatz eines photosensitiven Linkers [154] beschrieben, der eine kovalente Bindung der Aptamere an die Zielmoleküle forciert. Für spezielle Zielmoleküle wie beispielsweise Transkriptionsfaktoren bietet sich darüber hinaus die Möglichkeit, von zufällig synthetisierten Bibliotheken auf solche überzugehen,

die aus einem konkreten Genom abgeleitet sind. Vorteil dieses Ansatzes ist das natürliche Vorkommen bindender Regionen im Genom, die bereits in der evolutionären Entwicklung optimiert wurden [155].

Die zum Aufbau der Bibliothek eingesetzten Primersequenzen werden häufig kritisch betrachtet, da sie auf die Prozesse während der Selektion Einfluss nehmen können. Kurze *Insert*-Regionen sind von diesem Effekt besonders betroffen, sodass nach Möglichkeiten gesucht wurde, diesen zu verringern. Im sogenannten *Tailored SELEX* [156] werden die eigentlichen Primersequenzen nur während der notwendigen Phasen der Amplifikation und Konditionierung an die Oligonukleotide angebracht, vor der Bindung jedoch wieder entfernt. Die dafür notwendigen Hybridisierungsstellen sind wesentlich kürzer als die Primer selbst und beeinflussen daher die Faltung und Bindung der Aptamere nur minimal. Mit der vollständigen Entfernung und aufwändigeren Rekonstruktion der Primersequenzen geht der Ansatz des *Primer Free SELEX* [157; 158] noch einen Schritt weiter [141].

Weitere Modifikationen Zahlreiche modifizierte SELEX-Verfahren grenzen sich über die Verwendung spezieller Methoden in den einzelnen Prozessschritten ab [148]. Vereinzelt wird jedoch auch mit den Paradigmen gebrochen. So finden sich Varianten ohne Amplifikation [159], mit nur einer Runde [91] oder mit interner Parallelisierung [160]. Schließlich spielt auch die Automatisierung [161] der Schritte eine Rolle.

3 Auswertung der Primär- und Sekundärstruktur von Nukleinsäuren

Der Prozess der Aptamerentwicklung umfasst neben der eigentlichen Selektion mit dem Verfahren SELEX in der Regel nur die Sequenzierung weniger, in der finalen Bibliothek hoch angereicherter Aptamerkandidaten. Obwohl dieses Vorgehen für die reine Aptamergewinnung zweckmäßig ist, erlaubt es keinerlei Rückschlüsse auf die Entwicklung der eingesetzten Bibliothek während der Selektion und deren genaue Zusammensetzung nach der letzten Runde. Beides kann jedoch für die Entwicklung leistungsfähigerer Aptamere einen Mehrwert bieten, da sich durch die Ausdifferenzierung bindender von nicht-bindenden Aptamerkandidaten wichtige Bindungsinformationen in der Bibliothek niederschlagen. Auf deren Basis kann die bioinformatische Analyse schließlich nicht nur Teilsequenzen identifizieren, die die Bindung zwischen Aptamer und Zielmolekül vermitteln, sondern auch abschätzen, wie sich diese graduell auf die Affinität der Aptamere auswirken. Die Datengrundlage für eine solche Analyse kann durch Verfahren der Hochdurchsatzsequenzierung wie NGS mit vertretbarem zeitlichen und finanziellen Aufwand geschaffen werden, muss jedoch für die numerischen Verfahren der Analyse geeignet zugänglich gemacht werden. Da die numerische Repräsentation der Nukleinsäuresequenzen essentiell für die bioinformatische Analyse ist, widmet sich dieses Kapitel der vergleichenden Evaluation derartiger Beschreibungskonzepte.

3.1 Numerische Beschreibung von Nukleinsäuren

Die numerische Beschreibung von Nukleinsäuren ist eine entscheidende Voraussetzung für die weitere analytische Aufarbeitung der im Experiment gewonnenen Daten, da sie eine Brücke zwischen dem konkreten biologischen Sachverhalt und der numerischen Zugänglichkeit seiner Daten für die Algorithmen schlägt. Für die numerische Beschreibung können aus den Nukleinsäuren zahlreiche unterschiedliche Kennwerte abgeleitet werden, die im weiteren die Bezeichnung Deskriptoren tragen. Diese umfassen neben statistischen und topologischen Eigenschaften der Sequenz auch unterschiedliche physikochemische Aspekte. Ziel ist dabei die akkurate und effektive Abbildung der biologischen Realität in die Ebene der mathematischen Abstraktion, was die Anwendung der Algorithmen auf biologische Daten ermöglicht, obwohl diese ausschließlich für die Verarbeitung numerischer Eingaben ausgelegt sind. Zur Generierung verlässlicher Erkenntnisse muss jedoch eine gute Abdeckung des komplexen biologischen Parameterraums durch die ausgewählten Deskriptoren sichergestellt werden. Bei der Zusammenstellung von Deskriptoren können sowohl experimentell bestimmte als auch auf Basis der Sequenz- und Strukturinformation *in silico* abgeleitete Kennwerte einbezogen werden. Da die Verfügbarkeit experimentell bestimmter Kennwerte sehr eingeschränkt ist, bilden diese für die bioinformatische Analyse nur eine unzureichende informationelle Basis, von der aus kein verlässliches Analyseergebnis erwartet werden kann. Numerische Deskriptoren, die aus der Sequenz und

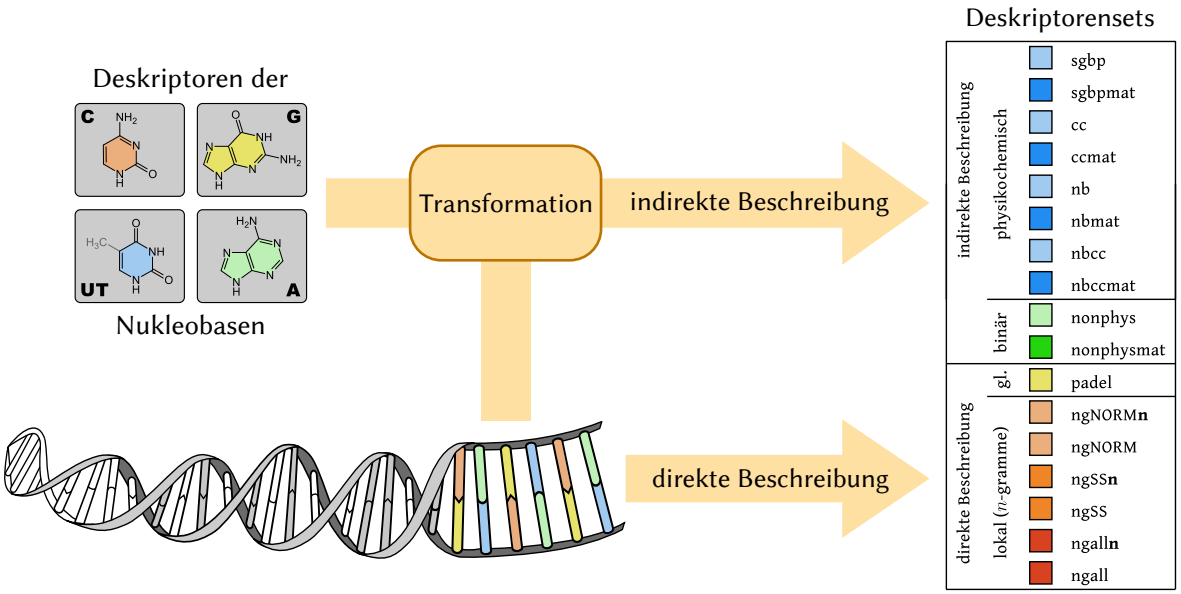


Abb. 3.1: Schematische Übersicht über die grundlegenden Beschreibungskonzepte für Nukleinsäuresequenzen und die konkret daraus abgeleiteten Deskriptorensets. Die indirekte Beschreibung kombiniert die Ableitung von Nukleobasendeskriptoren mit Transformationsstrategien, die deren Anwendung auf ganze Nukleinsäuresequenzen erlauben. Es wird in physikochemische (blau hervorgehoben) und binäre (grün hervorgehoben) Deskriptorensets unterschieden. Die direkte Beschreibung umfasst Ansätze, die unmittelbar auf die Nukleinsäuresequenz angewendet werden. Sie werden unterteilt in globale (gl., gelb hervorgehoben) und lokale Deskriptoren (orange und rot hervorgehoben).

Struktur der Nukleinsäuren abgeleitet werden, können hingegen als zuverlässige und leicht zugängliche Informationsquelle dienen. Sie gewinnen damit große Relevanz für die bioinformatische Analyse von Nukleinsäuren.

In der Literatur finden sich verschiedene Konzepte zur Beschreibung von Nukleinsäuren anhand ihrer Sequenz und Struktur. Diese stützen sich hauptsächlich auf zwei Arten von Information, die physikochemische und die positionelle Information. Einige Ansätze beruhen auf der Tatsache, dass jede der fünf Nukleobasen eine jeweils für sie charakteristische, physikochemische Signatur aufweist und die Kombination dieser Signaturen in der späteren Nukleinsäure verantwortlich für die Ausprägung ihrer eigentlichen Funktion ist. Die daraus resultierende zweiteilige Beschreibungsarchitektur umfasst die Ableitung von Nukleobasendeskriptoren auf Grundlage verschiedener physikochemischer Aspekte [162–164] und die Anwendung dieser auf ganze Nukleinsäuresequenzen mithilfe von speziellen Transformationsverfahren [165–170]. Ferner entwickelten sich auch Verfahren, die eine direkte Beschreibung der Nukleinsäuren ohne intermediäre Nukleobasendeskriptoren anstrebten [171–173].

Der Aufbau dieses Abschnittes orientiert sich an dem schematischen Überblick in Abbildung 3.1. Er beginnt mit der Vorstellung von Nukleobasendeskriptoren unterschiedlicher Kategorien, welche von den folgend eingeführten Transformationsstrategien auf Nukleinsäuremoleküle angewendet werden können. Die Kombination dieser beiden Ansätze wird sowohl in der Abbildung als auch im weiteren Verlauf des Kapitels als indirekte Beschreibung der Nukleinsäuren bezeichnet. Schließlich wird das Augenmerk auf die Konzepte zur direkten Beschreibung von Nukleinsäuren gelenkt, welche den Zwischenschritt über die Nukleobasendeskriptoren nicht erfordern.

3.1.1 Nukleobasendeskriptoren

Die Kombination der spezifischen physikochemischen Eigenschaften der einzelnen Nukleobasen bildet zusammen mit deren individuellen Ausrichtungen im Raum entlang des makromolekularen Stranges eine komplexe und einzigartig geprägte Oberfläche. Da mit dieser Oberfläche die Funktion der Nukleinsäure einhergeht, wurde in verschiedenen Ansätzen versucht, numerische Deskriptoren für Nukleobasen abzuleiten. Die Ansätze konzentrieren sich dabei auf verschiedene physikochemische Aspekte der Nukleobasen [162–164]. Für die Anwendung dieser Deskriptoren auf Nukleinsäuren muss jedoch der Weg über eine geeignete Transformation gegangen werden.

Scores of Generalized Base Properties Ein Ansatz zur Beschreibung der Nukleobasen sind die sogenannten *Scores of Generalized Base Properties* (SGBP), die für jede der fünf Nukleobasen physikochemische Informationen aus insgesamt 1209 molekularen Einzeldeskriptoren beziehen. Die Deskriptoren stammen aus vier Hauptgruppen und werden im folgenden entsprechend als 0D- bis 3D-Deskriptoren bezeichnet [162]. Einfache Deskriptoren, welche weder von der molekularen Gesamtkonformation noch von der atomaren Vernetzung abhängen, werden in der Gruppe der 0D-Deskriptoren zusammengefasst. Das sind zum Beispiel Zählattribute über verschiedene Atom- und Bindungsarten und aufsummierte physikochemische Eigenschaften wie das Molekulargewicht [162; 174]. Zur zweiten Gruppe, den 1D-Deskriptoren, zählen diejenigen einfachen Deskriptoren, in welche die molekulare Gesamtkonformation oder die atomare Vernetzung einfließen. Hierzu zählen unter anderem Zählattribute über funktionelle Gruppen [162; 174; 175]. Komplexere Deskriptoren, die sich hauptsächlich aus dem molekularen Graphen ableiten, gehören zur dritten Gruppe, den 2D-Deskriptoren. Diese beinhalten topologische und autokorrelationsbasierte Deskriptoren und weitere Indizes, welche sich beispielsweise aus der Konnektivität, dem Informationsgehalt und der Kantenadjazenz ableiten. In dieser Gruppe werden zudem topologischen Invarianten und Eigenwerte zur Beschreibung genutzt [162; 176–181]. Die letzte Gruppe der 3D-Deskriptoren umfasst diejenigen Kennwerte, die Informationen aus den räumlichen Atomkoordinaten und der molekularen Geometrie beziehen. Das schließt auch daraus abgeleitete Maße wie interatomare Abstände und Distanzverteilungen ein, setzt jedoch eine exakte dreidimensionale Molekülstruktur voraus. Zahlreiche der 3D-Deskriptoren kombinieren dabei die Verarbeitung der Rauminformationen mit einer Gewichtung auf Basis atomarer Eigenschaften [162; 182–188]. Zur Berechnung der zugrundeliegenden 1209 Einzeldeskriptoren wurden von den Autoren verschiedene Softwarelösungen, unter anderem Dragon [171] und GITMHDV [189], verwendet. Die Kombination unterschiedlicher Tools und Konzepte bedingte jedoch eine hohe Redundanz innerhalb dieses Beschreibungsdatensatzes, die durch das Entfernen aller Einzeldeskriptoren mit hoher paarweiser Korrelation untereinander reduziert wurde. Die verbliebenen 41 geringfügig korrelierten Deskriptoren dienten als Ausgangspunkt für eine Hauptkomponentenanalyse, deren vier erste Hauptkomponenten den Großteil der Eingabevarianz abdeckten ohne jedoch untereinander eine Korrelation zu zeigen [162]. Diese vier Hauptkomponenten bilden die Nukleobasendeskriptoren der SGBP und sind in Tabelle 3.1a zu finden.

***Non-Bonded Interaction*- und *Charge-Charge Interaction*-Deskriptoren** Zwei weitere Ansätze konzentrieren sich auf physikochemische Parameter, die in definierten Positionen der räumlichen Umgebung der einzelnen Nukleobasen durch Simulation bestimmt wurden. Die Autoren bezogen in ihre Untersuchung neben den fünf Standardnukleobasen auch zwölf

weitere, chemisch modifizierte Nukleobasenvarianten ein [163]. Die Qualität der zugrundeliegenden Molekülstrukturen ist für eine solche Simulation ein unerlässliches Kriterium, welches sicherstellt, dass die Verwendung der Ergebnisse in korrekten und aussagekräftigen Deskriptoren resultiert. Nach einer eingehenden experimentellen und theoretischen Untersuchung der fünf Standardnukleobasen [190] wurden die daraus resultierenden Molekülstrukturen als Startkonfiguration eingesetzt. Für die zwölf chemisch modifizierten Nukleobasenvarianten wurden mithilfe der Software MODEL[191] entsprechende Startkonfigurationen erzeugt und schließlich in der *Molecular Orbital Package* (MOPAC) [192] Softwaresuite einer semiempirischen Optimierung unter Anwendung des *Austin Model 1* (AM1)[193] unterzogen [163]. Die resultierenden 17 Strukturen wurden im Anschluss so zueinander ausgerichtet, dass das aufgespannte Gitternetz bei der späteren Parameterextraktion vergleichbare Werte liefert. Dabei wurde der Koordinatenursprung an die Position des Atoms gelegt, welches die Verbindung zum Zucker herstellt. Der im Modell durch ein Wasserstoffatom ersetzte Zucker wurde auf der negativen Ordinatenachse fixiert. Die Ringstrukturen der Nukleobasen wurden ferner auf der X-Y-Ebene des Koordinatensystems zueinander ausgerichtet [163]. Schließlich legten die Autoren ein festes, dreidimensionales Gitter mit einer Rasterung von $1{,}5\,\text{Å}$ um die ausgerichteten Strukturen und erhielten damit 1386 definierte Messpunkte.

Basierend auf dieser Vorbereitung wurden mit *Non-Bonded Interaction* (NBI) und *Charge-Charge Interaction* (CCI) zwei Gruppen von Nukleobasendeskriptoren auf der Grundlage unterschiedlicher physikochemischer Informationen abgeleitet. Durch simuliertes Abtasten der 17 Nukleobasenstrukturen mit einer Methylgruppe wurde im ersten Durchlauf an jedem dieser Gitterpunkte ein physikochemischer Parameter bestimmt, der die nichtkovalenten Bindungsspezifika der Nukleobasen, auch NBI, beschreibt. Aus den 1386 Einzelwerten des Parameters wurden schließlich nach einer Zentrierung um den Mittelwert durch eine Hauptkomponentenanalyse die fünf finalen NBI-Deskriptoren $\text{NBI}_1, \ldots, \text{NBI}_5$ abgeleitet [163]. Die gleiche Verfahrensweise wurde in einem zweiten Durchlauf mit veränderter physikochemischer Parametrisierung wiederholt. Hier wurden die elektrostatischen Wechselwirkungen der Nukleobase mit der Umgebung, auch CCI, durch Simulation einer Punktladung beschrieben und in die finalen CCI-Deskriptoren $\text{CCI}_1, \ldots, \text{CCI}_3$ überführt [163]. Beide Deskriptorensets finden sich in Tabelle 3.1b.

***Binary Coded Nucleobase Information*-Deskriptoren** Es ist unbestritten, dass der explizite Einsatz von physikochemischen Informationen den eben vorgestellten Deskriptoren eine charakteristische Prägung mit hoher Relevanz für die Beschreibung molekularer Interaktionen verleiht. Zur gültigen numerischen Beschreibung der Nukleobasen besteht jedoch keine Notwendigkeit zur Nutzung derartiger Zusatzinformationen. Als möglicher Vertreter einer solchen nicht-physikochemischen Beschreibung gilt die binäre Kodierung der Nukleobasen. Diese im Folgenden als *Binary Coded Nucleobase Information* (BCNI) bezeichnete Kodierung nutzt insgesamt fünf Deskriptoren, welche entsprechend der fünf betrachteten Nukleobasen benannt sind. Wie in Tabelle 3.1c gezeigt wird, ist der Wert der BCNI für die jeweils namentlich korrespondierende Nukleobase mit dem Wert 1 und für alle weiteren mit dem Wert 0 belegt [194; 195]. Bezugnehmend auf dieses binäre Kodierungsverfahren resultiert die Anwendung gängiger Distanzmaße in gleichen Abständen zwischen den jeweiligen Nukleobasen. Diese künstliche Gleichsetzung war schon mehrfach Anlass von Kritik [163], wird jedoch in dieser Arbeit nicht als Problem gewertet, da unter Berücksichtigung der kleinen Zahl betrachteter Nukleobasen die Distanzen, die durch eine beliebige Kodierung ausgedrückt werden, nur gering ins Gewicht fallen.

Tab. 3.1: Übersicht über die verschiedenen numerischen Deskriptoren für Nukleobasen, welche sowohl explizite (a, b) als auch implizite (c) physikochemische Information beinhalten.

(a) Die *Scores of Generalized Base Properties* [162] fassen eine breitgefächerte Vielzahl theoretischer molekularer Deskriptoren zusammen.

	Nukleobase				
Deskriptor	**A**	**C**	**G**	**T**	**U**
$SGBP_1$	-3,9505	4,3677	-2,7552	0,4217	1,9163
$SGBP_2$	4,0764	1,0541	-4,8467	0,8763	-1,1601
$SGBP_3$	-1,1507	1,5173	1,1540	3,3983	-4,9190
$SGBP_4$	1,2426	3,2084	1,4321	-4,0915	-1,7917

(b) Die *Non-Bonded Interaction-* und *Charge-Charge Interaction-*Deskriptoren [163] basieren auf der 3D-Simulation zweier physikochemischer Parameter.

	Nukleobase				
Deskriptor	**A**	**C**	**G**	**T**	**U**
NBI_1	73,05	-95,49	132,34	-133,31	-81,33
NBI_2	-27,34	-63,55	4,14	89,60	-42,99
NBI_3	-43,97	-9,62	24,74	-1,58	6,23
NBI_4	-35,86	4,79	12,14	11,74	7,70
NBI_5	40,93	10,06	16,26	-16,66	-29,04
CCI_1	-36,28	-161,02	162,40	-18,06	-9,26
CCI_2	63,42	19,53	35,09	-75,64	-82,95
CCI_3	12,07	-17,74	-22,03	-37,59	-42,28

(c) Die *Binary Coded Nucleobase Information*-Deskriptoren nutzen eine einfache binäre Kodierung, welche keinerlei explizite Anreicherung von physikochemischen Informationen enthält.

	Nukleobase				
Deskriptor	**A**	**C**	**G**	**T**	**U**
NP_A	1	0	0	0	0
NP_C	0	1	0	0	0
NP_G	0	0	1	0	0
NP_T	0	0	0	1	0
NP_U	0	0	0	0	1

3.1.2 Transformationsstrategien

Um die Vernetzung der einzelnen Nukleobasen durch das makromolekulare Rückgrat und die sekundärstrukturellen Kontakte beschreiben zu können, ist es notwendig geeignete Transformationsstrategien anzuwenden. Die in der Literatur beschriebenen Ansätze nutzen sowohl absolute als auch relative Positionsinformation zur Kodierung der einzelnen Nukleobasendeskriptoren. Die folgende Auswahl repräsentiert die grundlegenden Konzepte.

längenabhängige Transformation Im einfachsten Ansatz werden die einzelnen Deskriptorenwerte für eine Nukleotidsequenz entsprechend der vorgegebenen sequenziellen Reihenfolge bestimmt und zu einem geordneten Vektor kumuliert. Diese Transformation liefert einen Beschreibungsvektor, dessen Größe von der Länge der beschriebenen Sequenz abhängt. Da die Algorithmen, die später mit der numerischen Beschreibung der Nukleinsäuren arbeiten, in der Regel auf Beschreibungsvektoren von konstanter Länge beschränkt sind, kann diese Transformation nur bei Sequenzen gleicher Länge angewendet werden. Das Verfahren wird daher im weiteren Verlauf des Kapitels als längenabhängige Transformation (LAT) bezeichnet. Sei eine Menge von m Deskriptoren gegeben und ferner für jeden Deskriptor $p \in [1,m]$ die zugehörige Abfolge von Deskriptorwerten durch ${}^{p}w = (w_{p1}, w_{p2}, \ldots, w_{pn})$ definiert, so wird die primäre $m \times n$ Deskriptormatrix $\mathbb{M}$ entsprechend der Formel 3.1 gebildet. Der gesuchte Beschreibungsvektor D wird schließlich wie in Formel 3.2 gezeigt durch die Vektorisierung der primären Deskriptormatrix gewonnen und enthält $m \cdot n$ Werte. Zusätzlich zum Informationsgehalt der zugrundeliegenden Nukleobasendeskriptoren beinhaltet der Beschreibungsvektor D absolute Positionsinformation, ermöglicht also die genaue positionelle Zuordnung der einzelnen Werte.

$$\mathbb{M} = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix} \tag{3.1}$$

$$\mathrm{D} = (w_{11}, \ldots, w_{1n}, \ldots, w_{m1}, \ldots, w_{mn}) \tag{3.2}$$

Autokorrelation Neben der LAT gibt es eine weitere Gruppe von Transformationen, welche sich bei der Kodierung der einzelnen Deskriptorenwerte dem Prinzip der Autokorrelation (AK) bedienen. Diese Ansätze berücksichtigen potentielle Abhängigkeiten benachbarter Nukleobasen entlang der Sequenz ohne jedoch dabei auf absolute Positionen der einzelnen Nukleobasen angewiesen zu sein. Die Nachbarschaftseffekte werden dabei in Form eines Abstandsparameters k in die numerische Transformation eingebunden, was sich in der Generierung relativer Positionsinformation manifestiert. Der Parameter k gibt dabei den Abstand der Nukleobasen an, deren Interaktion als relevant erachtet wird. Ist eine Sequenz von Deskriptorwerten $w = (w_1, w_2, \ldots, w_n)$ gegeben, so können sowohl die Autokorrelationen nach Moreau-Broto ATS_k in normaler (siehe Formel 3.3, [165]) und gemittelter Form $\overline{\mathrm{ATS}}_k$ (siehe Formel 3.4, [166]), als auch die Koeffizienten nach Moran I_k (siehe Formel 3.5, [167]) und Geary c_k (siehe Formel 3.6, [168]) nach einem ähnlichen Prinzip berechnet werden. Wird statt der Summenbildung lediglich das erreichbare Maximum bei Multiplikation der Deskriptorenwerte der entsprechend voneinander entfernten Paare von Nukleobasen betrachtet, so beschreibt das Ergebnis die einfache maximale Autokorrelation MAC_k (siehe Formel 3.7) der Sequenz von Deskriptorwerten [174]. Zur Auswahl der relevanten Interaktionen findet dabei das Kronecker-Delta $\delta_{ij}^{k} = \delta(d_{ij}; k)$ Anwendung, welches für alle Paare von Nukleobasen mit einer topologischen Distanz von k den Wert 1 annimmt und

für alle weiteren 0. Die topologische Information für die Berechnung des Kronecker-Deltas kann dabei entweder aus der einfachen Sequenznachbarschaft oder aus der Sekundärstruktur der Nukleinsäure abgeleitet werden. Δ_k gibt dabei an, wie viele Paare ij eine topologische Distanz von k aufweisen und kann als Doppelsumme $\Delta_k = \sum_{i=1}^{n}\sum_{j=i}^{n}\delta_{ij}^{k}$ berechnet werden. Schließlich werden die angeführten autokorrelationsbasierten Transformationen mit einem definierten Wertebereich für den Parameter k auf alle Sequenzen von Deskriptorwerten angewendet, die sich aus einem Set von Nukleobasendeskriptoren ergeben. Die daraus resultierenden Werte werden schließlich zum finalen Beschreibungsvektor D zusammengefasst.

$$\mathrm{ATS}_k = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i \cdot w_j \cdot \delta_{ij}^{k} \tag{3.3}$$

$$\overline{\mathrm{ATS}}_k = \frac{\mathrm{ATS}_k}{\Delta_k} \tag{3.4}$$

$$I_k = \frac{n}{\Delta_k} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(w_i - \bar{w})(w_j - \bar{w})\delta_{ij}^{k}}{\sum_{i=1}^{n}(w_i - \bar{w})^2} \tag{3.5}$$

$$c_k = \frac{n-1}{2\Delta_k} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(w_i - w_j)^2\delta_{ij}^{k}}{\sum_{i=1}^{n}(w_i - \bar{w})^2} \tag{3.6}$$

$$\mathrm{MAC}_k = \max_{i=1}^{n}\max_{j=1}^{n}\left(w_i \cdot w_j \cdot \delta_{ij}^{k}\right) \tag{3.7}$$

Kreuzautokorrelation Die Berechnungsvorschriften der eben vorgestellten AKs-basierten Verfahren erlauben eine Modifikation, mit deren Hilfe es möglich wird, zwei unterschiedliche Nukleobasendeskriptoren in die Berechnung einfließen zu lassen. Das Ergebnis enthält dann ebenfalls relative Positionsinformationen der beschriebenen Nukleinsäuren. Sind zwei Sequenzen von Deskriptorwerten ${}^{1}w = ({}^{1}w_1, {}^{1}w_2, \ldots, {}^{1}w_n)$ und ${}^{2}w = ({}^{2}w_1, {}^{2}w_2, \ldots, {}^{2}w_n)$ gegeben, so erhält man die korrespondierende Kreuzautokorrelation (KK), indem das erste Auftreten des jeweiligen Deskriptorwertes w_i durch den der ersten Sequenz ${}^{1}w_i$ und das zweite Auftreten des Deskriptorwertes w_j durch den der zweiten Sequenz ${}^{2}w_j$ ersetzt wird. Die entsprechend der Formeln 3.8 bis 3.12 berechneten KK-Deskriptoren beschreiben Nukleinsäuren unter Zuhilfenahme zweier unterschiedlicher Nukleobasendeskriptoren. Der finale Beschreibungsvektor D wird analog zu dem der AKs-basierten Verfahren gebildet.

$$\mathrm{ATS}_k^{\times} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} {}^{1}w_i \cdot {}^{2}w_j \cdot \delta_{ij}^{k} \tag{3.8}$$

$$\overline{\mathrm{ATS}}_k^{\times} = \frac{\mathrm{ATS}_k^{\times}}{\Delta_k} \tag{3.9}$$

$$I_k^{\times} = \frac{n}{\Delta_k} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}({}^{1}w_i - \overline{{}^{1}w})({}^{2}w_i - \overline{{}^{2}w})\delta_{ij}^{k}}{\sum_{i=1}^{n}({}^{1}w_i - \overline{{}^{1}w})({}^{2}w_i - \overline{{}^{2}w})} \tag{3.10}$$

$$c_k^{\times} = \frac{n-1}{2\Delta_k} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} ({}^1w_i - {}^2w_j)^2 \delta_{ij}^k}{\sum_{i=1}^{n} ({}^1w_i - \overline{{}^1w})({}^2w_i - \overline{{}^2w})} \tag{3.11}$$

$$\mathrm{MAC}_k^{\times} = \max_{i=1}^{n} \max_{j=1}^{n} \left({}^1w_i \cdot {}^2w_j \cdot \delta_{ij}^k \right) \tag{3.12}$$

Lineare und bilineare Indizes Der Ansatz, der hinter den linearen und bilinearen Indizes steht, ist den AK- und KK-Transformationen im Prinzip sehr ähnlich, macht jedoch auf andere Weise Gebrauch von den zur Verfügung stehenden topologischen Informationen der Nukleinsäuren [174]. Sind zwei Sequenzen von Deskriptorwerten ${}^1w = ({}^1w_1, {}^1w_2, \ldots, {}^1w_n)$ und ${}^2w = ({}^2w_1, {}^2w_2, \ldots, {}^2w_n)$ zusammen mit einer Adjazenzmatrix $\mathbb{M}$ gegeben, so können die Indizes berechnet werden. Diese graphentheoretische Matrix $\mathbb{M}$ enthält dabei die topologischen Informationen, die entweder der primären Sequenzkonnektivität oder der Sekundärstruktur entnommen werden. Durch die Anwendung der k-ten Potenz auf die Adjazenzmatrix $\mathbb{M}$ wird ein Effekt erzielt, der ähnlich dem des Parameters k der AK- und KK-Transformationen ist. Jedoch wird bei Verwendung von Potenzen der Adjazenzmatrix die Konnektivität zweier Positionen auf der Sequenz besonders bei Einbezug der Sekundärstrukturinformationen feiner aufgelöst. Die totalen, nicht-stochastischen linearen Indizes f_k (siehe Formel 3.13) nutzen dabei analog zu den AK nur eine Sequenz von Deskriptorwerten zur Berechnung. Die Kombination beider Sequenzen ist jedoch mit den totalen, nicht-stochastischen bilinearen Indizes b_k (siehe Formel 3.14) möglich. Für beide Arten existieren Normalisierungsstrategien, welche hier nicht betrachtet werden [169; 170].

$$f_k = \sum_{i=1}^{n}\sum_{j=1}^{n} \left[\mathbb{M}^k\right]_{ij} \cdot {}^1w_j \tag{3.13}$$

$$b_k = \sum_{i=1}^{n}\sum_{j=1}^{n} \left[\mathbb{M}^k\right]_{ij} \cdot {}^1w_i \cdot {}^2w_j \tag{3.14}$$

3.1.3 Direkt anwendbare Beschreibungskonzepte

Eine sachgemäße numerische Beschreibung von Nukleinsäuren muss aber nicht zwingend den Umweg über die initiale Charakterisierung der einzelnen Nukleobasen und die folgende Anwendung von Transformationsfunktionen einschlagen. Vielmehr existieren auch Verfahren, die eine direkte Beschreibung der Nukleinsäuren ermöglichen. Die nachfolgend vorgestellte Auswahl an Verfahren setzt dabei ebenfalls auf relative und absolute Positionsinformationen unterschiedlicher molekularer Ebenen.

Worthäufigkeiten Ein ursprünglich aus der Sprachverarbeitung stammender, zeichenkettenbasierter Ansatz nutzt die Auftretenshäufigkeiten einzelner Worte als positionsunabhängige Beschreibungsmerkmale des Eingabetextes. Da es jedoch für die Anwendung an biologischen Sequenzen kein geeignetes Äquivalent für menschliche Worte gibt [173], werden im diesem Kontext gemeinhin feste Wortlängen für die Ableitung der Merkmale vorgegeben [196]. Die Festlegung der Wortlänge ist ein kritischer Punkt in der Vorbereitung der Beschreibung, da die Aussagekraft bei kurzen Worten sehr schnell abnimmt, die Anzahl möglicher Deskriptoren jedoch exponentiell mit der Wortlänge steigt und daher die Weiterverarbeitung der erzeugten Beschreibung behindert. Während es in geschriebenen Texten nicht zur Überschneidung von

Worten kommt, muss eine solche bei der Betrachtung biologischer Sequenzen in Kauf genommen werden. Da keine Informationen darüber existieren, welcher Teil der umliegenden Sequenz tatsächlich für die weitere Verarbeitung relevant ist, wird die Sequenz mit einem beweglichen Fenster durchlaufen [196]. Die gezählten Fragmente überschneiden sich somit entsprechend ihrer sequenziellen Nachbarschaft. Sie stellen dank ihrer Unabhängigkeit von der absoluten Sequenzposition eine wichtige Quelle relativer Positionsinformationen dar, die im Vergleich zu den folgenden Deskriptoren tatsächlich lokal in der Sequenz verortet werden kann. Die hieraus resultierenden Deskriptoren werden abhängig von der gewählten Wortlänge n als n-Gramme bezeichnet. Durch die Erweiterung des Grundalphabetes um modifizierte Nukleobasen-Zeichen wird das zusätzliche Einbringen der Sekundärstrukturinformationen (SSI) der betrachteten Nukleinsäure möglich. Mit Hinblick auf eine gute Lesbarkeit für Menschen bietet es sich an, im erweiterten Alphabet unverpaarte Nukleobasen durch Kleinbuchstaben und verpaarte durch Großbuchstaben zu kennzeichnen.

Molekulare Deskriptoren auf atomarer Ebene Wenn die atomare Struktur eines Moleküls bekannt ist, so steht eine große Zahl von molekularen Deskriptoren zur Verfügung, die direkt auf diese Struktur angewendet werden können. Das schließt nicht nur einfache Häufigkeitskennwerte über Atome, Bindungen und funktionelle Gruppen ein, sondern auch komplexere strukturelle und graphenbasierte Deskriptoren, welche die 2D-Geometrie der Moleküle einbeziehen. Zu den genannten zählen beispielsweise fragmentbasierte Ansätze wie *Substructural Molecular Fragments* (SMF) und *ISIDA Property-Labelled Fragments* (IPLF). Deskriptoren werden aber auch aus Interaktionsfeldern abgeleitet, wie es beispielsweise bei *3D Molecular Interaction Field* (MIF) und *Grid-independent molecular Descriptors* (GRIND) der Fall ist. Darüber hinaus kommen auch Technologien wie die AK- und KK-Transformationen auf atomarer Ebene zur Anwendung. Wenn die erforderlichen atomaren Strukturen weder vorhanden sind, noch experimentell erzeugt werden können, dann besteht unter Anwendung geeigneter Software die Möglichkeit der computergesteuerten Generierung einer Prototypenstruktur [197; 198]. Eine solche ist nach ihrer Erzeugung weder korrekt gefaltet, noch befinden die einzelnen Atome in einer verlässlichen räumlichen Ausrichtung. In den meisten Fällen liegen die Prototypstrukturen in einer strukturell linearisierten oder anhand der Sekundärstruktur bereits teilweise vorgefalteten Form vor. Diese spiegelt jedoch die korrekte atomare Topologie des Moleküls sowie im lokalen Umfeld auch eine gute Näherung der relativen Lage naher Atome zueinander wider. Mit Ausnahme einiger molekularer Deskriptoren, die die exakte dreidimensionale Geometrie der Struktur erfordern, ist dies als Grundlage für die Berechnung ausreichend.

Durch die große Zahl verfügbarer molekularer Deskriptoren für die atomare Ebene ist deren manuelle Berechnung für alle Eingabestrukturen praktisch nicht sinnvoll. Aus diesem Anlass entstanden mit der Zeit Softwarelösungen, die dem Benutzer das automatische Berechnen verschiedener Teilgruppen dieser Deskriptoren ermöglichten. Einige bekannte Vertreter derartiger Softwarelösungen sind in Tabelle 3.2 aufgeführt. Unter diesen sollen besonders die zwei Tools Dragon [171] und PaDEL-Descriptor [203] hervorgehoben werden. Beide stellen dem Nutzer jeweils eine universelle Sammlung von Berechnungsroutinen zur Verfügung und wurden bereits erfolgreich in Studien verwendet, um *Quantitative Structure-Activity Relationship* (QSAR)-Modelle zu generieren [208; 209]. Während Dragon ein kommerzielles Produkt ist, handelt es sich bei PaDEL um ein kostenfrei verfügbares *Open Source*-Produkt. Die molekularen Deskriptoren auf atomarer Ebene wurden in einer Zeit entwickelt, in der sich die Analyse von Struktur-Wirkungsbeziehungen und damit auch die Notwendigkeit der numerischen Beschreibung auf

Tab. 3.2: Übersicht verschiedener Softwareprodukte zur Berechnung molekularer Deskriptoren auf Basis atomarer Strukturen. Mit einem Stern (*) gekennzeichnete Produkte werden kommerziell vermarktet. Stand Februar 2016.

Name	Beschreibung	Referenz
Dragon*	universell	[171]
Codessa Pro*	universell	[199]
Corina Symphony*	universell	[200]
Molecular Operating Environment*	universell	[201]
Chemistry Development Kit (CDK)	universell	[202]
PaDEL-Descriptor	universell	[203]
Pentacle*	GRIND	[204]
VolSurf+*	MIF	[205]
ISIDA Fragmentor	SMF und IPLF	[206]
AFGen	fragmentbasiert	[207]

kleine Moleküle beschränkte. Bedingt durch das Missverhältnis aus algorithmischer Zielstellung und Größenordnung der eingegebenen Moleküle kann die Anwendung der Deskriptoren auf Makromoleküle zu verschiedenen Problemen führen. Neben dem Erreichen der Berechnungsgrenzen einiger Deskriptoren sind an dieser Stelle besonders die strukturellen Defizite zu benennen. Durch die Zielstellung kleiner Moleküle ist der Fokus der Deskriptoren in den meisten Fällen auf spezifische Charakteristika der atomaren Ebene festgelegt. Diese sind zwar in makromolekularen Strukturen ebenfalls enthalten und wirksam, nehmen hier jedoch nur eine untergeordnete Rolle ein.

3.2 Konzeption einer Strategie zur Evaluation der Beschreibungskonzepte

Aufgrund der schlechten Datenlage wurde die Evaluation nicht anhand von Aptamersequenzen, sondern auf Basis eines Datensatzes von Promotoren als alternativen funktionellen Nukleinsäuren durchgeführt. Der Datensatz beinhaltete Informationen über die individuellen Wirkungsstärken der einzelnen Promotoren als Referenzgröße. Die anhand der Beschreibungskonzepte abgeleiteten Deskriptoren wurden zu Gruppen zusammengestellt und zur numerischen Beschreibung auf die Promotoren des Datensatzes angewendet. Anschließend wurde vergleichend untersucht, wie genau die bekannten Wirkungsstärken der Promotoren durch ein geeignetes Regressionsmodell vorhergesagt werden konnten. Da im untersuchten System sowohl das mathematische Regressionsverfahren als auch der verwendete Datensatz unverändert blieben, konnten Änderungen in der Güte des Vorhersageergebnisses dem zugehörigen Deskriptorenset zugeschrieben werden. Das Wirkungsprinzip der Promotoren gilt als hinreichend erforscht und verstanden, sodass die Ergebnisse der jeweiligen Modellierung vor diesem Hintergrund verifiziert werden konnten. Die durchgeführte Evaluation förderte durch die hinreichend systematisierte Betrachtung außerdem das Verständnis über die Charakteristika der einzelnen Beschreibungskonzepte und ließ Rückschlüsse auf die Wichtigkeit der informationelle Komponenten zu.

3.2.1 Beschreibung und Vorverarbeitung des Datensatzes

Für die Durchführung der Evaluation konnte kein hinreichend großer Datensatz von Aptamersequenzen zu einem gemeinsamen Zielmolekül gefunden werden, indem alle Kandidaten experimentell in Bezug auf ihre Affinität charakterisiert wurden. Aufgrund der schlechten Datenlage musste daher auf andere Vertreter der funktionellen Nukleinsäuren ausgewichen werden. Dieses Vorgehen war zulässig, da die verschiedenen Arten funktioneller Nukleinsäuren auf ein ähnliches Wirkungsprinzip aufbauen. Nicht nur die Funktion der Aptamere beruht demnach auf der molekularen Erkennung des Zielmoleküls, auch die Funktion von Promotoren und regulatorischen RNAs kann auf dieses Prinzip zurückgeführt werden. So tritt die DNA eines Promotors mit der RNA-Polymerase in Interaktion, um den sogenannten offenen Komplex zu formen, der ein wichtige Rolle im Transkriptionsprozess einnimmt [210]. Kleine, regulatorische RNAs binden nach dem Komplementaritätsprinzip an stark konservierte Argonautenproteine und steuern damit die sequenzspezifische Ausschaltung von Messenger-RNAs [211]. In all diesen Fällen sind verschiedene Formen von Bindungsmotiven an der charakteristischen Funktion der jeweiligen Nukleinsäure maßgeblich beteiligt. Für die Evaluation der Deskriptoren konnte daher stellvertretend ein Datensatz von Promotorsequenzen genutzt werden.

Biologischer Hintergrund Promotoren sind wichtige Regelungselemente im Prozess der Genexpression. Sie sind selbst nicht kodierend und befinden sich auf der Gensequenz in der Nähe des Transkriptionsstartpunkts am $5'$-Ende kodierender Bereiche. Durch die Interaktion von Promotor und RNA-Polymerase entsteht der offene Komplex, im dem die DNA-Doppelhelix kurzzeitig zum Ablesen entwunden wird. Der Promotor ist damit essentiell für die Gentranskription [210]. Besonders bakterielle Promotoren zeichnen sich durch eine sehr einheitliche Struktur aus. Dabei wird die Interaktion mit der Polymerase hauptsächlich von zwei charakteristischen Sequenzmotiven des Promotors vermittelt. Das erste Motiv mit der Konsensussequenz `TTGACA` liegt in der -35-Region des Promotors und wird auch als *RNA Polymerase Site* (RPS) bezeichnet. Das zweite Motiv, die sogenannte *Pribnow Box*, liegt in der -10-Region und weist die Konsensussequenz `TATAAT` auf. Beide Motive werden in der Literatur durchgehend als essentiell wichtig für die Funktion und bestimmend für die Leistungsfähigkeit von Promotoren beschrieben, die im folgenden auch als ihre Wirkungsstärke bezeichnet wird [212–214]. Die Klassifikation von Gensequenzfragmenten als Promotoren und die Relation zwischen Promotorsequenz und der verbundenen Wirkungsstärke wurden bereits in einer Serie von Publikationen behandelt [162; 164; 194; 195; 215; 216].

Vorstellung des Datensatzes In dieser Untersuchung wird ein Datensatz von 38 Promotorsequenzen verwendet, welcher sowohl originale Promotoren von *E. coli* als auch solche enthält, die durch eine Infektion von λ-Bakteriophagen verändert wurden [194; 195; 215; 217; 218]. Dieser Datensatz wurde bereits in der Vergangenheit verwendet, um die Anwendbarkeit und den Nutzen eines physikochemischen Deskriptorensets für Nukleobasen zu bekräftigen [162]. Die einzelnen Promotorsequenzen, welche eine Länge von $68\,\mathrm{nt}$ aufweisen, umfassen den Bereich von $-49\,\mathrm{nt}$ bis $19\,\mathrm{nt}$ relativ zum Transkriptionsstart. Sie umschließen damit die RPS und die *Pribnow Box*, also die beiden Sequenzmotive, welche sich in der Vergangenheit als besonders wichtig für die Aktivität der Promotoren und die erreichbare Transkriptionsgeschwindigkeit herausgestellt haben [162; 212–214]. Die Promotoren aus Tabelle 3.3 wurden von den Autoren hinsichtlich ihrer relativen Wirkstärke in Bezug auf einen internen Standard charakterisiert, der durch den Referenzpromotor für β-Laktamase vorgegeben war [162; 195; 219; 220]. Die

Tab. 3.3: Promotorendatensatz für die Evaluation der Beschreibungskonzepte für Nukleinsäuresequenzen. Der Datensatz umfasst Namen, Sequenzen und logarithmische, relative Promotorwirkstärken (logRS) von 38 DNA-Promotoren [162]. Die Sequenzen sind entsprechend der jeweiligen Nukleobasen eingefärbt. Dabei weisen Pastelltöne auf reguläre Nukleobasen hin. Intensive Farbtöne markieren diejenigen Nukleobasen, welche am Aufbau der beiden wichtigen Sequenzmotive beteiligt sind. Die vollständigen Motive (*RNA Polymerase Site* und *Pribnow Box*) sind als Referenz unter der Tabelle entsprechend angeordnet.

Name	Sequenz	logRS
D/E20	ACTGCAAAAATAGTTTGACACCCTAGCCGATAGGCTTTAAGATGTACCCAGTTCGATGAGAGCGATAA	1.748
H207	TTAAAAAATTCATTTGCTAAACGCTTCAAATTCTCGTATAATATACTTCATAAATTGATAAACAAAAA	1.740
G25	GAAAAATAAAATTCTTGATAAAATTTTCCAATACTATTATAATATTGTTATTAAAGAGGAGAAATTAA	1.278
A1	ATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCATCGAGAGGGACACGGCGA	1.881
2	GAAAAACAGGTATTGACAACATGAAGTAACATGCAGTAAGATACAAATGCCTAGGTAACACTAGCAGC	1.301
L	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAGGACGCACTGAC	1.568
CON	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACCATAAGGAGGTGGATCCGGC	0.602
LAC	AGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTGGTATGTTGTGTGGAATTGTGAGCGGATAACAA	0.756
LAC/UV5	AGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTGGTATAATGTGTGGAATTGTGAGCGGATAACAA	0.518
N25/03	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCAATTGTGAGCGGATAACAA	0.903
N25/ANTI	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATCCGGAATCCTCTTCCCG	0.432
N25/LAC	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCAAATTGTGAGCGGATAACA	0.903
CON/N25	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATAAATTTGAGAGAGGAGT	1.398
CON/PEX	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATAAAGGGTCGAGAGGAGT	1.204
CON/ANTI	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATCCGGAATCCTCTTCCCG	0.255
CON/D/E20	TTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACCCAGTTCGATGAGAGCGATAA	1.114
CON/TRP	TTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACGCAAGTTCACGTAAAAAGGGT	0.903
L-8A	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATAATGAGCACATCAGCAGGACGCACTGAC	1.672
L/CON	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATCAGCAGGACGCACTGAC	1.146
L/N25	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATAAATTTGAGAGAGGAGT	1.813
L/CON/N25	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATAAATTTGAGAGAGGAGT	1.813
N25/05	GGATAACAATTTAGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.173
N25/USR	GGCTAAAAAACACGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.491
CON/05	GGATAACAATTTAGTTGACATTTTTAAGCTTGGCGGTTATAATGTTACCATAAGGAGGTGGGAATTCC	1.173
L/N25DSR	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATAAATTTGAGAGAGGAGT	1.813
L/CON/N25DSR	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATAAATTTGAGAGAGGAGT	1.778
LS1	TCCGTCTCGACGGGTTGACACAAAAGCCACAAGGGGTTATAATGAGCACATAAACTTGAGAGAGGAAT	2.143
LS2	TGCGTATAGACAGTTTGACACAAAAGCCACAAGGTGTTATAATGAGCACATAAATTTGAGAGAGGAAT	2.217
N25	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.477
J5	TATAAAAACCGTTATTGACACAGGTGGAAATTTAGAATATACTGTTAGTAAACCTAATGGATCGACCT	0.954
A3	TGAAACAAAACGGTTGACAACATGAAGTAAACACGGTACGATGTACCACATGAAACGACAGTGAGTCA	1.342
TACL	TTCTGAAATGAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAA	1.230
N25/PEX	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAAGGGTCGAGAAGAGT	1.176
CON/03	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCAATTGTGAGCGGATAACAA	0.903
L-12T	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATACTGAGCACATCAGCAGGACGCACTGAC	1.398
N25/04	CATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTTATAATAGATTCATAAATTTGAGAGAGGAGT	1.246
N25/AUSR	GGCTGTGCGGCACGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.301
L/N25USR	CATAAAAAATTTATTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAGGACGCACTGAC	1.763
Motive	TTGACA TATAAT	

angewendete Charakterisierungsstrategie wird sowohl gegenüber translatorischen Effekten als auch gegenüber Veränderungen in der Zahl der Genkopien als wertneutral beschrieben. Da in der bestimmten Wirkungsstärke große Variationen gemessen wurden, wurde der Messwert zur weiteren Verwendung logarithmiert. Große Werte stehen dabei für eine hohe Wirkstärke. Die resultierende logarithmische, relative Promotorstärke (logRS) wurde als Zielgröße für die Optimierung der Regressionsmodelle und die darauffolgende Beurteilung der Deskriptorensets eingesetzt [162; 195].

Durchgeführte Vorverarbeitung Da zu den jeweiligen Promotorsequenzen keine entsprechenden 3D-Strukturinformationen vorlagen, wurde die Software 3DNA [198] zur Erzeugung von Prototypstrukturen verwendet, auf deren Basis die molekularen Deskriptoren der atomaren Ebene abgeleitet werden konnten. Einige Studien wiesen darauf hin, dass die Genexpression auch über andere, nicht ausschließlich sequenzspezifische Modi der genetischen Kodierung reguliert werden kann. So wurde ein Zusammenhang zwischen der Aktivität und Sekundärstrukturausprägung bei Promotoren beobachtet [221; 222]. Um diese zusätzliche Informationsquelle in der Betrachtung nutzen zu können, wurden die SSI bei der Erzeugung der n-Gramme durch die Einführung zusätzlicher Symbole einbezogen. Zur Vorhersage von Sekundärstrukturen wurde das Softwarepaket ViennaRNA [223] genutzt.

3.2.2 Zusammenstellung von Deskriptorensets

Die Verwendung einzelner Deskriptoren ist nicht hinreichend für die Beschreibung der Nukleinsäuren. Die vorgestellten Konzepte wurden daher kombiniert und entsprechend den zugrundeliegenden Prinzipien zu Gruppen, den sogenannten Deskriptorensets, zusammengegliedert. Die Sets stellten die Grundlage für die Evaluation der Beschreibungskonzepte dar.

Indirekte Beschreibung In einem ersten Schritt wurden die verschiedenen Nukleobasendeskriptoren mit den vorgestellten Transformationen kombiniert, um Deskriptorensets zur indirekten Beschreibung der Nukleinsäuren zu erhalten. Zur besseren Differenzierung der absoluten und relativen Positionsinformationen wurden aus den Nukleobasendeskriptoren (SGBP, CCI, NBI, NBI+CCI und BCNI) unter Anwendung unterschiedlicher Transformationen jeweils genau zwei Deskriptorensets erzeugt. Das erste Set ergab sich aus der Anwendung der LAT, das zweite vereinigt die AK- und KK-Transformationen mit den linearen und bilinearen Indices und lieferte so eine längenunabhängige Transformation (LUT) der entsprechenden Nukleobasendeskriptoren. Der Abstandsparameter wurde dabei auf einen Wertebereich von $1 \leq k \leq 8$ beschränkt.

Direkte Beschreibung Im zweiten Schritt wurden die Sets für die direkte Beschreibung zusammengestellt. Von den aufgeführten Softwarelösungen zur Erzeugung molekularer Deskriptoren wurden die kommerziellen Produkte ausgeschlossen, um die Nachvollziehbarkeit der Arbeit auch ohne finanziellen Aufwand zu ermöglichen. Ferner wurde auf den Einsatz derjenigen Vertreter verzichtet, die nur eine geringe Vielfalt und Anzahl an Deskriptoren bereitstellten. Da PaDEL die CDK-Bibliothek verwendet [203], muss diese nicht separat betrachtet werden. In der Kategorie der direkten, globalen Beschreibung blieben daher nur die PaDEL-Deskriptoren zur weiteren Verwendung. Aufgrund der Parametrisierbarkeit der n-Gramm-Deskriptoren fiel die Entscheidung hier auf das Anlegen mehrerer Deskriptorensets mit unterschiedlicher Wortlänge und Sekundärstrukturbeteiligung. Dies erleichterte bei Eignung die spätere Auswahl der optimalen Parameter. Für Fragmentlängen von eins bis vier wurde jeweils ein Set ohne SSI, ein Set mit SSI und das Vereinigungsset angelegt. Um auch die Kombination verschiedener Fragmentlängen betrachten zu können, existiert für diese drei Kategorien jeweils noch ein Set mit Fragmenten aller Längen.

Systematisierung Aus den vorgestellten Beschreibungskonzepten für Nukleinsäuren wurden insgesamt 26 Deskriptorensets für die weitere Verwendung abgeleitet. Tabelle 3.4 liefert eine systematisch aufgebaute Zusammenfassung der Sets, welche sie in verschiedene relevante Kategorien einordnet und ein Farbschema zur einfacheren Wiedererkennung im Verlauf der Arbeit vorgibt.

3.2.3 Eingesetzte Methoden

Die korrekte numerische Repräsentation der physikochemischen und topologischen Charakteristika, die für die Interaktion zwischen Nukleinsäure und Bindepartner, hier Promotor und Polymerase, verantwortlich sind, ist von essentieller Bedeutung. Sie hilft bei der Erzeugung robuster Vorhersage- und Klassifikationsmodelle, die später zur Charakterisierung unbekannter Sequenzen genutzt werden können. Nach diesem Prinzip wurde die Güte der numerischen Repräsentation aus der Vorhersagegenauigkeit damit trainierter Regressionsmodelle bestimmt.

Tab. 3.4: Systematische Übersicht über die zusammengestellten Deskriptorensets. Die indirekte Beschreibung durch Nukleobasendeskriptoren resultierte in längenunabhängig transformierten (heller Farbton) und längenabhängig transformierten Deskriptoren (dunkler Farbton), welche sich im oberen Teil der Tabelle befinden. Sets mit explizit eingebrachten physikochemischen Informationen (blau) werden dabei getrennt von denen aufgeführt, die ohne diese zusätzliche Informationsquelle auskommen (grün). Im unteren Teil findet sich die direkte Beschreibung, welche hinsichtlich ihrer Fähigkeit, für lokale Strukturelemente angemessen Rechnung zu tragen, in globale (gelb) und lokale (orange, rot) Beschreibung unterteilt wird. Die n-Gramm-Deskriptoren werden dabei in einer gekürzten Weise dargestellt. Die Kennzeichnung $(\times 4)$ gibt dabei an, dass der Eintrag den vier Deskriptorensets entspricht, die durch Einsetzen der Fragmentlänge ($1 \leq \mathbf{n} \leq 4$) für den Parameter $\mathbf{n}$ im Namen erhalten werden.

		Setbezeichnung	Enthaltene Deskriptoren	
indirekte Beschreibung	physikochemisch	sgbp	SGBP in längenunabhängiger Transformation	
		sgbpmat	SGBP in längenabhängiger Trans.	
		cc	CCI-Deskriptoren in längenunabhängiger Trans.	
		ccmat	CCI-Deskriptoren in längenabhängiger Trans.	
		nb	NBI-Deskriptoren in längenunabhängiger Trans.	
		nbmat	NBI-Deskriptoren in längenabhängiger Trans.	
		nbcc	NBI+CCI-Deskriptoren in längenunabhängiger Trans.	
		nbccmat	NBI+CCI-Deskriptoren in längenabhängiger Trans.	
	binär	nonphys	BCNI-Deskriptoren in längenunabhängiger Trans.	
		nonphysmat	BCNI-Deskriptoren in längenabhängiger Trans.	
direkte Beschreibung	gl.	padel	PaDEL-Deskriptoren	
	lokal (n-Gramme)	ngNORM**n**	$\boldsymbol{n}$-Gramme ohne SSI $(\times 4)$	$1 \leq \boldsymbol{n} \leq 4$
		ngNORM	alle n-Gramme ohne SSI	
		ngSS**n**	$\boldsymbol{n}$-Gramme mit SSI $(\times 4)$	
		ngSS	alle n-Gramme mit SSI	
		ngall**n**	$\boldsymbol{n}$-Gramme mit und ohne SSI $(\times 4)$	
		ngall	alle n-Gramme mit und ohne SSI	

Dazu wurden die 26 Deskriptorensets aus Tabelle 3.4 auf die 38 Promotorsequenzen aus Tabelle 3.3 angewendet, um numerische Eingaben für die folgende Regression zu erhalten. Für jede Promotorsequenz und jedes Deskriptorenset wurde hierfür ein Vektor von konkreten Deskriptorwerten erzeugt. Die resultierenden Vektoren für ein Deskriptorenset waren dabei zwischen den einzelnen Sequenzen des Datensatzes vergleichbar. Auch wenn nicht explizit mitgeführt, können die Komponenten der Vektoren eindeutig zu den korrespondierenden Deskriptoren zugeordnet werden. Vor der Weiterverarbeitung wurden in einem Filterschritt gegenstandslose Deskriptoren aus den Beschreibungssets entfernt, also solche, die für alle Promotorsequenzen denselben Wert annahmen oder vielfache anderer Deskriptoren waren.

Die Eignung der Deskriptoren wurde schließlich durch das Trainieren und Auswerten von Regressionsmodellen auf der Basis dieser Beschreibungssets bewertet. Als mathematische Abstraktion der Promotorfunktion wurde die relative Wirkungsstärke logRS zur Zielgröße der Regression gewählt. Da weder die sonstigen Regressionsparameter noch der zugrundeliegende

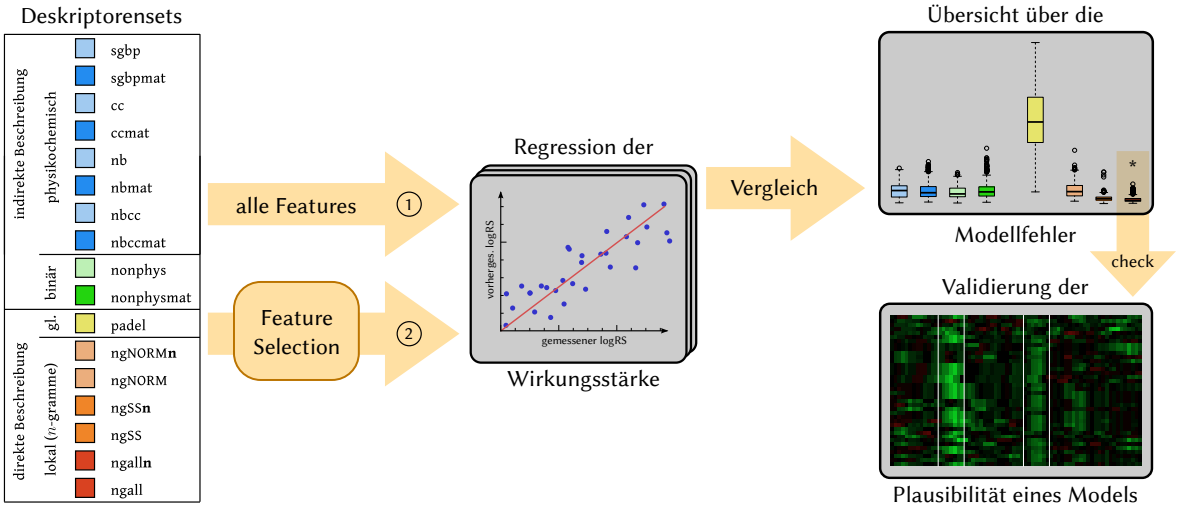


Abb. 3.2: Schematische Abbildung des Vergleichs- und Evaluationsprozesses. Der Vergleichsprozess beinhaltet die Regression sowohl mit (2) als auch ohne (1) *Feature Selection* und den eigentlichen Vergleich der Regressionsergebnisse. Zur Evaluation wird die Plausibilität des besten Modells (mit Stern markiert) geprüft.

Datensatz verändert wurde, kann die Regressionsgenauigkeit als Indikator für die Eignung der Deskriptorensets verwendet werden. Über den kreuzvalidierten *Root-Mean-Square Error* (RMSE) wurde auf die Regressionsgenauigkeit und schließlich auch auf die Eignung der Deskriptoren geschlossen. Dieser liefert Informationen über die Abweichung zwischen den gemessenen und den vom jeweiligen Regressionsmodell vorhergesagten Wirkungsstärken der Promotoren. Da kleine Werte für den RMSE dabei für wirklichkeitsgetreue Modelle stehen, deutet dies auf gute Eignung der eingesetzten Deskriptoren hin. Auch wenn einige Autoren von der Verwendung des RMSE als Maß der Modellgüte abraten [224; 225], wurde gezeigt, dass der RMSE in seiner Aussage nicht missverständlich ist und daher angewendet werden kann [226]. Um den Modellfehler möglichst frei von Einflüssen der Wahl von Trainings- und Testset zu bestimmen, wurden 50 randomisierte Wiederholungen einer zehnfachen Kreuzvalidierung (KV) durchgeführt. Entsprechend werden im Folgenden nicht nur einfache Mittelwerte betrachtet, sondern zusätzlich dazu auch die Streuungen in Form von Boxplots. Werte außerhalb des anderthalbfachen Interquartilsabstands werden im Boxplot als Ausreißer gekennzeichnet. Jedes Deskriptorenset enthält naturgemäß einen großen Anteil von Deskriptoren, die in der Modellgenerierung keinen Gewinn bringen. Da große Mengen solcher unzweckmäßiger Deskriptoren die Leistungsfähigkeit der Regressionsmodelle negativ beeinflussen können, wurde zur Verringerung dieser Störgröße zusätzlich eine *Feature Selection* durchgeführt.

Wie in Abbildung 3.2 skizziert, wurden zwei Durchläufe der Regression für jedes Beschreibungsset durchgeführt. Im ersten wurden alle verfügbaren Deskriptoren des Sets zur Regression eingesetzt, um einen groben Einblick in die Eigenschaften der unterschiedlichen Beschreibungskonzepte zu bekommen. Im zweiten Durchlauf wurde die *Feature Selection* angewendet, um den Einblick durch die verringerten Störfaktoren weiter zu vertiefen. Um die Verlässlichkeit der Erkenntnisse zu verbessern, stützte sich die Evaluation dabei hauptsächlich auf den zweiten Durchlauf, bei dessen Auswertung auch die statistische Signifikanz einfloss. Das Hauptergebnis der Evaluation stellt eine Rangfolge der Deskriptorensets dar, die auf deren Fähigkeit basiert, die hinter der Wirkstärke der Promotoren liegenden biologischen Prinzipien bestmöglich numerisch abzubilden. Darauf aufbauend wurde schließlich auf die Einflüsse der physikochemischen und positionellen Information bei der Beschreibung der Nukleinsäuren geschlossen.

Um die Zuverlässigkeit der Ergebnisse sicherzustellen, wurde im Anschluss an den Vergleich das Regressionsmodell, welches als das beste befunden wurde, auf Plausibilität hin überprüft. Dafür wurden die Beiträge der einzelnen Nukleobasen am Vorhersageergebnis über die Promotorensequenzen aufgetragen und in Relation zu den bekannten biologischen Charakteristiken der Promotoren gestellt. Hierbei wurden die Informationen zu den beiden wichtigen Sequenzmotiven, also deren Vorkommen, Ausformung und Position als Grundlage genutzt.

Regressionsverfahren Bei der Auswahl eines Regressionsverfahrens aus der Vielzahl der verfügbaren linearen und nicht-linearen Vertreter musste eine zweckdienliche Balance zwischen Modellkomplexität und Interpretierbarkeit gefunden werden. Darüber hinaus gab es im vorliegenden Fall einige Nebenbedingungen für die Anwendbarkeit der Modelle auf die gegebenen Daten zu beachten. So sollte das Verfahren der großen Anzahl an Deskriptoren und deren untereinander korreliertem Wesen geschuldet in der Lage sein, mit der Multikollinearität der Daten umzugehen. In den meisten multivariaten Methoden stellt Multikollinearität zwar keine generelle Verletzung der Grundannahmen dar, kann jedoch ernsthafte Probleme bei der Parameterschätzung hervorrufen [227; 228]. Unter zusätzlicher Berücksichtigung der geringen Anzahl an verfügbaren Sequenzen im Datensatz erfordern die meisten multivariaten Methoden vor ihrer Anwendung wenigstens eine exzessive Vorverarbeitung und Dimensionsreduktion der Daten.

Das lineare Regressionsverfahren *Partial Least Squares* (PLS)-Regression hat sich in diesem Zusammenhang als effektives und effizientes Werkzeug herausgestellt, welches problemlos in der Lage ist, die beiden Anforderungen des Datensatzes handzuhaben [229–232]. Besonders im wissenschaftlichen Kontext der Bioinformatik, Chemometrie und Pharmakologie hat die PLS-Regression in den letzten Jahrzehnten eine große Aufmerksamkeit erlangt [229; 232–234]. Sobald jedoch nicht-lineare Zusammenhänge zwischen den Daten bestehen, sind zur optimalen Repräsentation dieser Zusammenhänge entsprechend nicht-lineare Verfahren notwendig [235–237], die wiederum eine komplexere Modellparametrisierung erfordern [238]. Die individuelle Optimierung dieser Parameter würde jedoch die Vergleichbarkeit der resultierenden Regressionsmodelle stark in Mitleidenschaft ziehen. Mit Hinblick auf die Erzeugung vergleichbarer Modelle wurde daher trotz möglicher nicht-linearer Zusammenhänge auf das lineare PLS-basierte Regressionsverfahren zurückgegriffen, auch wenn dies eine möglicherweise geringere Vorhersagegenauigkeit nach sich zog.

PLS ist in der Lage, ein lineares Regressionsmodell für hochkorrelierte Datensätze durch das iterative Aufdecken latenter Zusammenhänge unter den korrelierten Variablen zu erzeugen. Das ist selbst dann noch möglich, wenn die Zahl der Variablen die Anzahl der Trainingsbeispiele des Datensatzes deutlich übersteigt [230; 239]. Die PLS-Regression ist in der Lage, einen entsprechenden Prädiktor $\hat{Y} = f(\mathbb{M}) \approx Y$ für jeden Datensatz von n Proben und m Variablen zu finden, der durch eine $n \times m$ Matrix $\mathbb{M}$ und einen Antwortvektor Y beschrieben werden kann. Dem Verfahren muss dabei mit der Anzahl der zu nutzenden latenten Variablen lediglich ein Parameter mitgeteilt werden. In der Literatur wird empfohlen, die Eingabedaten vor der Regression so zu transformieren, dass jede Eingabevariable eine Standardabweichung von 1 hat [240]. Dies ist der Tatsache zuzuschreiben, dass nichtstandardisierte Variablen durch die unterschiedlichen Ausmaße der jeweiligen Wertebereiche ungewollt gewichtet in die Regression eingehen. Die Standardisierung beugt diesem Effekt vor. Die Anwendung eines Standardisierungsfilters auf die Matrix $\mathbb{M}$ bewirkte folglich eine leichte Verbesserung der Vorhersagegenauigkeit. Ein erzeugtes Regressionsmodell kann durch einen Vektor β von Regressionskoeffizienten beschrie-

ben werden, der in Bezug auf den Einfluss der einzelnen Variablen sehr gut interpretierbar ist. Dies ist dem linearen Zusammenhang geschuldet, der im Modell zur Beschreibung konstruiert und genutzt wird. In dieser Arbeit wurde die PLS-Implementierung des *Waikato Environment for Knowledge Analysis* (Weka)-Frameworks [241; 242] genutzt.

Feature Selection Zur Verringerung des Regressionsfehlers wurde eine genetische *Feature Selection* angewendet, welche als Heuristik eine nahe-optimale Teilmenge von *Features* bestimmt [243–245]. Inspiriert durch die Prinzipien der biologischen Evolution und der natürlichen Selektion wurden durch das Verfahren verschiedene zufällig zusammengestellte Teilmengen von Eingabevariablen, den Deskriptoren, als Individuen einer initialen Population betrachtet. In diesem Kontext wird ein Individuum durch einen Bitstring beschrieben, der festlegt, welche der Deskriptoren des zugrundeliegenden Beschreibungssets dem Individuum angehören. Durch die Bewertung der Teilmenge mithilfe eines PLS-Regressionsmodells wurde für jedes Individuum ein individueller Fitnesswert bestimmt. Diejenigen Individuen mit hoher Fitness überleben den evolutionären Prozess mit höherer Wahrscheinlichkeit, da sie überlegene Modelle repräsentieren. Das System fördert somit den Erhalt leistungsstarker Teilmengen der eingegebenen Beschreibungssets. Zu den Prinzipien, die aus der Natur adaptiert wurden, zählen zudem die zufällige Mutation und die Rekombination zweier Elternindividuen. Die notwendigen Elternindividuen wurden nach dem Tournierverfahren als beste von jeweils vier zufällig ausgewählten Individuen bestimmt. Es wurde eine Java-Implementierung des genetischen Algorithmus entwickelt, die mit dem Weka-Framework [241; 242] kompatibel ist.

Der genetische Algorithmus kultivierte eine Population von 500 Individuen, welche zufällig mit einer Beladung von $10\,\%$ initialisiert wurden. Nach der direkten Übernahme der besten 25 Individuen in die Folgepopulation wurde der feste Anteil von $80\,\%$ der Folgepopulation durch Rekombination generiert und der Rest durch punktmutierte Individuen aufgefüllt. Dabei wurde eine Mutationsrate von $0{,}01$ angenommen. Der genetische Algorithmus terminierte, sobald die Verbesserungsrate der Populationsfittness unter einen bestimmten Schwellwert gefallen war, jedoch frühestens nach 100 Generationen. Durch diesen Ansatz wurden unnötige Deskriptoren, die sich vormals negativ auf die Vorhersagequalität der Regression auswirkten, aus den Beschreibungssets entfernt. Auf diese Weise wurde einerseits die Aussagekraft der entnehmbaren Bewertungen erhöht und andererseits die Möglichkeit geschaffen, wichtige Einflussfaktoren feingranularer zu betrachten.

Statistische Evaluierung Die aus den Regressionsmodellen durch KV abgeleiteten RMSE-Werte wurden als Maß für die Eignung der Deskriptoren zur Beschreibung von Promotoren als Stellvertreter von funktionellen Nukleinsäuren verstanden. Kleinere Modellfehler wiesen dabei auf Deskriptorensets hin, welche die festgelegte biologische Referenzgröße logRS am besten abbilden konnten. Beim Einsatz der RMSE-Werte zu Bewertungs- und Auswahlzwecken werden primär die auftretenden Werteunterschiede betrachtet. Entsprechend wurde die Evaluierung der statistischen Signifikanz dieser Wertunterschiede zur weiteren Fundierung der gewonnenen Erkenntnisse verwendet.

Die statistische Signifikanz von gemessenen Wertedifferenzen kann durch eine Reihe von bewährten statistische Tests beurteilt werden. Liegen normalverteilte Vergleichsgrößen vor, so kann der Welch-Test [246] verwendet werden, um die Signifikanz der auftretenden Unterschiede in den Mittelwerten zu beurteilen. Es handelt sich dabei um eine modifizierte Form des t-Tests nach Student [247], welcher in der Lage ist, mit ungleichen Varianzen in den beiden Ver-

gleichsdatensätzen umzugehen. Eine Überprüfung der vorliegenden RMSE-Daten ergab, dass die jeweiligen Datensätze nicht normalverteilt waren, sondern eine erhebliche Schiefe in der Verteilung der RMSE-Werte aufwiesen. Da diese jedoch gegen die Bestimmungen des Welch-Tests verstößt, wird generell auf die Verwendung nicht-parametrischer Statistik verwiesen. Der bekannteste Signifikanztest aus dieser Gruppe ist der Wilcoxon-Mann-Whitney-Test [248], welcher die Eingabedaten über eine interne Transformation in eine Rangliste überführt, um diese Einschränkung zu überwinden.

Sowohl der *t*-Test nach Student als auch der darauf aufbauende Welch-Test sind in der Tat am besten für normalverteilte Größen geeignet. Entgegen der verbreiteten Ansicht sind diese jedoch unempfindlich gegenüber Abweichungen von der Normalität, wenn die zu bewertenden Größen durch hinreichend viele Messungen beschrieben werden [249]. In diesem Zusammenhang wurde gezeigt, dass der *t*-Test nach Student sogar gegenüber stark verzerrten Verteilungen robust ist, wenn der Stichprobenumfang eine Größe von 200 übersteigt [250]. Selbiges kann auch für den Welch-Test angenommen werden. Obwohl der Wilcoxon-Mann-Whitney-Test keinerlei Anforderungen an die Verteilung selbst hat, hängt die Verlässlichkeit seiner Ergebnisse stark von der Homoskedastizität der zwei zu vergleichenden Proben ab, also der Gleichheit ihrer Verteilungen in Form und Varianz [249]. Er wird daher als besonders nützlich für Größen der Ordinalskala erachtet, kann aber unter bestimmten Bedingungen auch in Studien mit kleinerem Stichprobenumfang eingesetzt werden, um die Signifikanz von Unterschieden in Mittelwerten und Medianen zu bewerten [249–251]. Bei Untersuchungen von Mittelwerten kontinuierlicher Variablen mit größerem Stichprobenumfang wird jedoch dringend angeraten, einen parametrischen Test wie den *t*-Test oder den Welch-Test anzuwenden. Dieser lieferte in der vergleichenden Studie von Fagerland ein naturgemäßeres Ergebnis, welches dem erwarteten Verhalten in Bezug auf die Zurückweisung nicht-signifikanter Unterschiede besser entsprach [250]. Da der Stichprobenumfang von 500 in dieser Arbeit hinreichend groß war und es sich beim RMSE um eine kontinuierliche Größe handelt, wurde für die Beurteilung der Signifikanz der Welch-Test gewählt. Auf diese Weise wurde neben den unterschiedlichen Varianzen auch der Heteroskedastizität und Schiefe der einzelnen Verteilungen entsprechend Rechnung getragen.

Visualisierung Nach dem Training eines PLS-Regressionsmodells besteht die Möglichkeit, die Regressionskoeffizienten, welche auch β-Faktoren genannt werden, auszuwerten. Über sie kann der Einfluss der einzelnen Eingangsvariablen auf die vorhergesagte Zielgröße ermittelt werden. Da die Eingabedaten vor Beginn der Regression einen Standardisierungsfilter durchlaufen haben, bezogen sich die β-Faktoren, die direkt vom Modell entnommen wurden, auf die standardisierten Eingabewerte. Um den Einfluss der Standardisierung auf die Regressionskoeffizienten zu kompensieren, war folglich ein rechnerischer Zwischenschritt notwendig. Dessen Ergebnisse, die korrigierten β^*-Faktoren, gaben nun die Einflüsse der eigentlichen Eingabevariablen in logRS-Einheiten wieder. Solange sich die verwendeten Deskriptoren einzelnen Nukleobasen zuordnen lassen, können diese direkt auf die Promotorsequenzen aufgetragen werden, um so einen schnellen Überblick darüber zu erhalten, wie sich die Beiträge zum Vorhersagemodell sequenziell über die Promotoren verteilen. Im Falle von n-Gramm-Deskriptoren müssen die vorliegenden Koeffizienten für n-Gramme zuerst mit den vorhandenen Sequenzen verrechnet werden, um diese Übersicht zu gewinnen. Zu diesem Zweck wurden alle Vorkommen eines jeden n-Gramms in den Sequenzen gesucht. Der dem jeweiligen n-Gramm zugeordnete β^*-Faktor wurde anschließend durch Addition des entsprechenden Bruchteils auf alle am n-Gramm-Fund beteiligten Nukleobasen gleichmäßig verteilt. Durch die Überlappungen angrenzender

n-Gramm-Funde setzten sich die Beiträge einiger Sequenzpositionen aus denen mehrerer β^*-Faktoren zusammen. Beim Lesen der entsprechenden Diagramme sollten diese Seiteneffekte beachtet werden.

Die so erhaltenen nukleobasenweisen Beiträge zum Vorhersagemodell wurden schließlich genutzt, um eine visuelle Repräsentation zu generieren. Da die Grundstruktur einer Liste von Sequenzen der einer Matrix entspricht, wurde die Heatmap als Form der Visualisierung gewählt. Sie bietet die Möglichkeit innerhalb einer rechteckig zweidimensionalen Datenstruktur die Größe eines definierten Kennwertes grafisch aufzutragen. Positive Beiträge zum Regressionsmodell wurden durch grüne und negative Beiträge durch rote Färbung signalisiert. Die Farbskala verläuft im neutralen Bereich über schwarz.

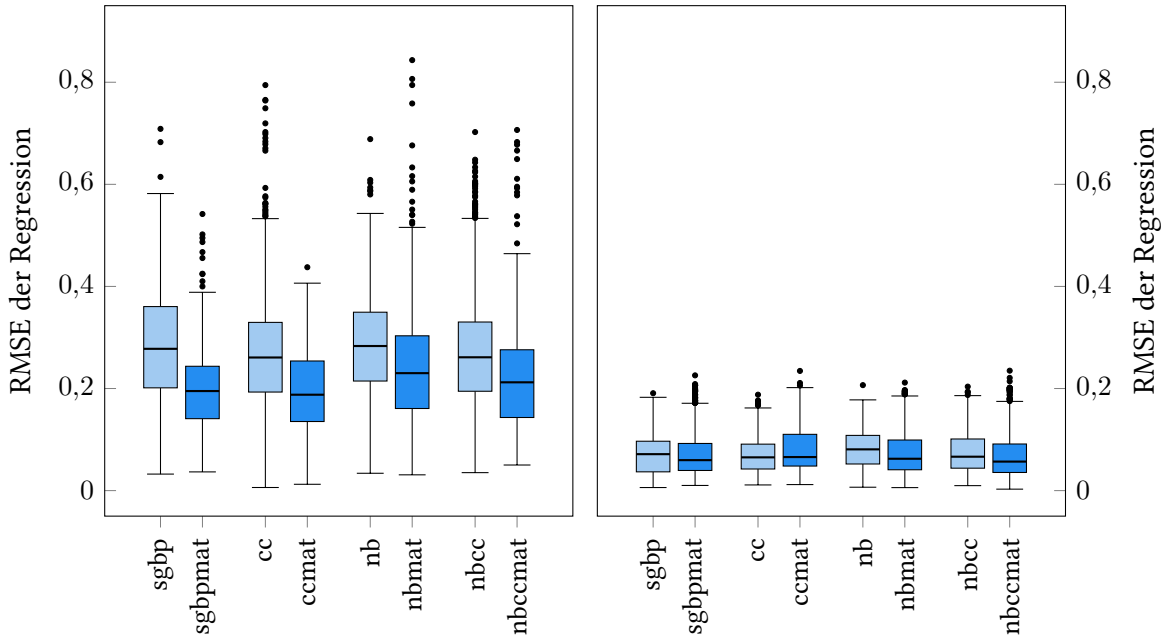
3.3 Ergebnisse der Evaluation

In diesem Abschnitt werden die Ergebnisse der Gegenüberstellung vorgestellt, interpretiert und diskutiert. Zusätzlich erfolgt eine Bewertung ihrer Plausibilität in Hinsicht auf die biologischen Hintergründe, sowie eine Überprüfung der Verhältnismäßigkeit der eingesetzten *Feature Selection*-Strategie. Der Abschnitt endet mit einer abschließenden, die Hauptergebnisse zusammenfassenden Betrachtung.

3.3.1 Gegenüberstellung der Beschreibungskonzepte

Die Gegenüberstellung der Beschreibungskonzepte erfolgt in themenbezogener Schrittweise. So werden in einer erster Instanz die Einflüsse sowohl der Transformationen der indirekten Beschreibung als auch der Anreicherung der Deskriptoren mit physikochemischer Information betrachtet. In einer zweiten Instanz werden dann die Einflüsse der Parametrisierung auf die Güte der n-Gramm-Deskriptoren untersucht. Schließlich werden die Ergebnisse der Betrachtungen zusammengeführt.

Einfluss der Transformationen der indirekten Beschreibung Während die längenabhängige Transformation die absolute Positionsinformation der Sequenz erhält, wird durch die eingesetzten längenunabhängigen Transformationen eine neue Form der Positionsinformation eingeführt. Diese relative Positionsinformation trägt der Korrelationen zwischen räumlich benachbarten Nukleobasen in definierten Abständen Rechnung. Auch wenn in beiden Fällen die physikochemische Information der unterliegenden Nukleobasendeskriptoren erhalten wird, ist anzunehmen, dass die unterschiedlichen Kodierungen dieser Information Einfluss auf die Verwertbarkeit in der Regression nehmen. Es empfiehlt sich daher, zunächst einen Überblick über den Einfluss der beiden Formen von Positionsinformation auf die Regressionsleistung zu gewinnen, um später gezielt eine der Transformationen auswählen zu können. Ohne eine Filterung der Deskriptorensets durch die *Feature Selection* ist durchgängig beobachtbar, dass auf Basis der LAT- im Vergleich zu den LUT-Deskriptoren Modelle mit kleineren RMSE-Werten erzeugt wurden. Siehe dazu Abbildung 3.3a. Wie jedoch in Abbildung 3.3b deutlich wird, führte die Auswahl wichtiger *Features* nicht nur zu einer allgemeinen Verringerung der Modellierungsfehler, sondern auch zu einer Veränderung der RMSE-Verhältnisse der beiden betrachteten Gruppen. Da die Unterschiede im Modellierungsfehler auf ein vernachlässigbaren Level abgefallen waren, kann der Einsatz der beiden Transformationen praktisch als gleichwertig erachtet werden.

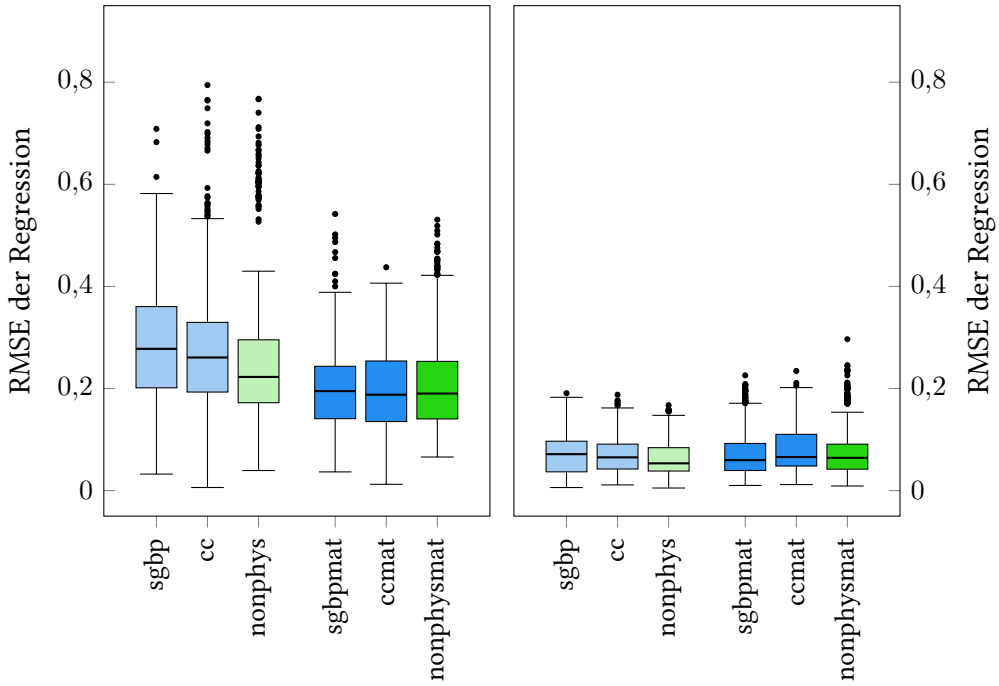


Verwendung aller Deskriptoren **(a)** **(b)** Nach *Feature Selection*

Abb. 3.3: Vergleich zwischen LAT (dunkles blau) und LUT (helles blau) auf Basis der explizit mit physikochemischen Informationen angereicherten Nukleobasendeskriptoren. Vergleichsgrößen sind die kreuzvalidierten RMSE-Werte der PLS-Regressionsmodelle, die auf den jeweiligen Beschreibungssets trainiert wurden. Die Farbgebung und Benennung der Beschreibungssets richtet sich nach den Vorgaben in Tabelle 3.4. Der deutliche Vorsprung der LAT in (a) ist nach der *Feature Selection* (b) nicht mehr vorhanden.

Liang *et al.* diskutierten in ihrer Studie die Anwendung von AK- und KK-Transformationen auf physikochemisch geprägte Nukleobasendeskriptoren und betonten dabei besonders den Nutzen bei Sequenzen unterschiedlicher Länge. Da die untersuchten Promotorsequenzen ihrer Studien jedoch alle die gleiche Länge aufwiesen, kamen die AK- und KK-Transformationen nicht zur Anwendung [162]. Auch wenn in der vorliegenden Arbeit gezeigt wurde, dass die Art der Positionsinformation nur einen geringen Einfluss ausübt und damit eine eher untergeordnete Rolle einnimmt, ließen die Beobachtungen eine schwach ausgeprägte Präferenz zur Nutzung der relativen Positionsinformation zu. Das deutet darauf hin, dass relative Abhängigkeiten innerhalb der Promotorsequenzen existieren, die von einer ausschließlichen Betrachtung der absoluten Sequenzpositionen nicht abgedeckt werden können, jedoch trotzdem in der Lage sind geringen Einfluss auf die Aktivität der Promotoren zu nehmen. Ohne *Feature Selection* zeigten die LUT jedoch eine Neigung zu höheren Modellfehlern, da durch die zahlreichen Deskriptorenkombinationen ein störendes Rauschen in die Daten kam. Ohne eine entsprechende Filterung führte dies zu einem negativen Einfluss auf das Regressionsmodell. Die *Feature Selection* entfernte diese störenden Effekte jedoch. Es kann also festgehalten werden, dass die Deskriptoren mit relativer denen mit absoluter Positionsinformation aus zwei Gründen vorgezogen werden, zum einen wegen der besseren Abdeckung relativer Abhängigkeiten innerhalb der Sequenzen und zum anderen wegen der Toleranz gegenüber unterschiedlich langen Sequenzen.

Einfluss der Anreicherung mit physikochemischer Information Nach der Untersuchung der verschiedenen Arten der Positionsinformation folgt nun die genauere Beleuchtung des praktischen Effekts, der durch das explizite Einführen von physikochemischer Information in die Nukleobasendeskriptoren erreicht wurde. Aus diesem Grund werden jeweils Deskriptoren mit



Verwendung aller Deskriptoren **(a)** **(b)** Nach *Feature Selection*

Abb. 3.4: Vergleich zwischen Deskriptoren mit (blau) und ohne (grün) expliziter Anreicherung physikochemischer Information, jeweils gruppiert nach Art der Transformation in LAT (dunkler Farbton) und LUT (heller Farbton). Vergleichsgrößen sind die kreuzvalidierten RMSE-Werte der PLS-Regressionsmodelle, die auf den jeweiligen Beschreibungssets trainiert wurden. Die Farbgebung und Benennung der Beschreibungssets richtet sich nach den Vorgaben in Tabelle 3.4. Sowohl vor als auch nach der *Feature Selection* gibt es nur minimale Unterschiede in der Aussagekraft der erzeugten Regressionsmodelle.

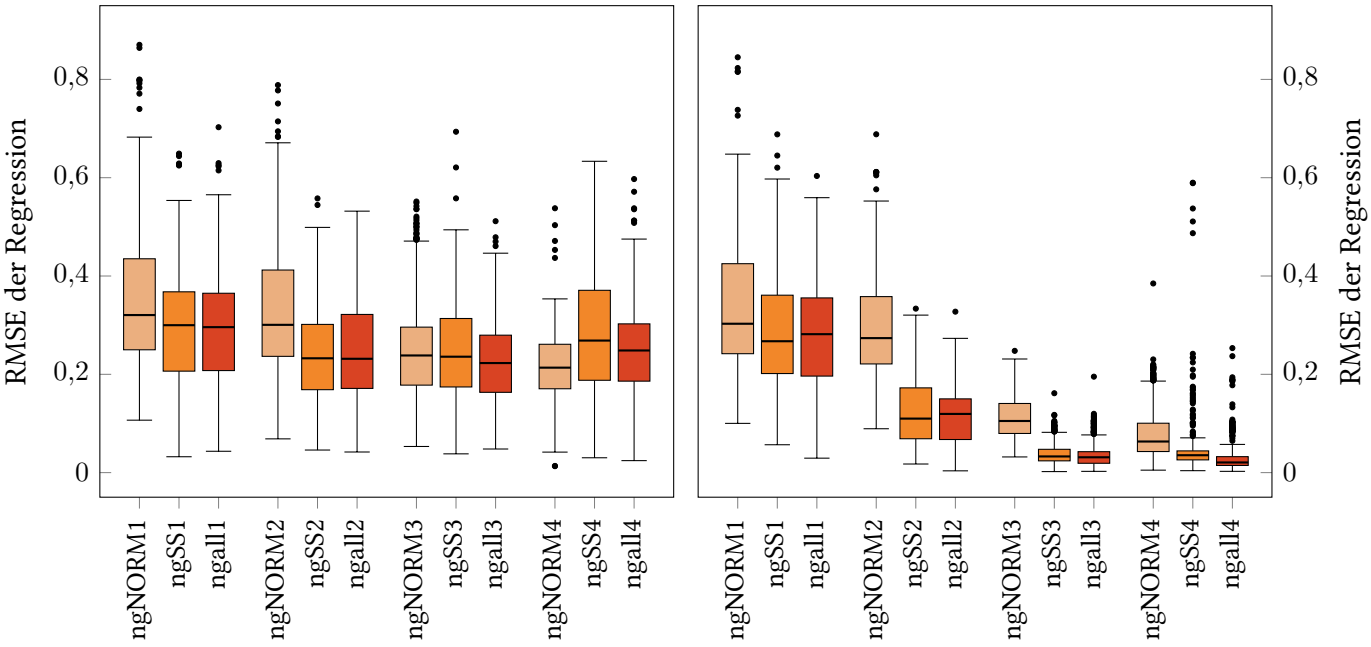
und ohne dieser expliziten physikochemischen Information gegenübergestellt. Die bereits vorgestellten BCNI-Deskriptoren wurden durch eine binäre Kodierung in Abwesenheit expliziter physikochemischer Zusatzinformationen erzeugt. Sie eignen sich daher als Vergleichspartner für die bereits untersuchten Beschreibungssets. Unter Einbezug aller Deskriptoren waren nur marginale Unterschiede zwischen den physikochemischen und den BCNI-Deskriptoren zu beobachten. Wie Abbildung 3.4a zeigt, lieferten die BCNI-Deskriptoren (`nonphys` und `nonphysmat`) unter den LUT-basierten Sets sogar geringere Modellierungsfehler und damit ein leicht besseres Ergebnis als die Deskriptoren auf explizit physikochemischer Basis. Nach der Auswahl wichtiger *Features* durch den genetischen Algorithmus zeigt der Vergleich in Abbildung 3.4b neben den allgemein niedrigeren RMSE-Werten keinen bemerkbaren Verlust an Vorhersagekraft der Modelle, der mit dem Wegfall der expliziten physikochemischen Informationen in den Deskriptoren in Verbindung gebracht werden konnte. Die Mittelwertunterschiede zwischen den längenabhängig transformierten Beschreibungssets `nonphysmat`, `ccmat` und `nbmat` wurden als nichtsignifikant ($p > 0.15$) bewertet. Bei den längenunabhängig transformierten Beschreibungssets ist eine leichte Verbesserung der Modellvorhersage durch das Weglassen der expliziten physikochemischen Information zu erkennen ($p < 0.01$). Dieser soll jedoch im weiteren Verlauf dieser Arbeit trotz der theoretischen Signifikanz aufgrund ihrer sehr geringen Ausprägung keine weitere Bedeutung zugemessen werden.

Im zweiten Schritt wurde überprüft, welchen Effekt der Wegfall der Positionsinformationen im Datensatz hervorruft. Das PaDEL-Deskriptorenset [203] enthielt zwar zahlreiche geometrische und topologische Deskriptoren und damit auch positionelle Informationen, diese waren

jedoch durch die Art der Deskriptoren auf die molekulare bzw. atomare Ebene beschränkt. Die positionelle Information lag daher in einer auf die Zielstellung bezogen inkompatiblen Größenordnung vor, sodass sie in Hinsicht auf die Beschreibung von Makromolekülen nur als geringfügig bis nicht nutzbringend betrachtet werden musste. Aus diesem Grund erfüllte das PaDEL-Deskriptorenset die Anforderung der Abwesenheit positioneller Informationen. Das Diagramm in Abbildung 3.6a zeigt, dass die Modelle, welche auf dem PaDEL-Beschreibungsset trainiert wurden, die höchsten RMSE-Werte aufwiesen und damit die schlechtesten Ergebnisse lieferten. Im Gegensatz zu allen anderen Beschreibungssets war beim PaDEL-Set keine signifikante Verbesserung der Modelle durch die Auswahl eines optimierten Subsets an *Features* erkennbar ($p > 0.55$). In Abbildung 3.6b wird daher deutlich, dass nach der *Feature Selection* der Unterschied zwischen den Ergebnissen des PaDEL-Sets und denen der anderen Beschreibungssets wesentlich größer ausfiel als noch davor.

Der Vergleich zwischen den physikochemisch angereicherten Nukleobasendeskriptoren und den BCNI hat keinen Verlust der Vorhersagekraft der Regressionsmodelle festgestellt, der sich mit dem Weglassen der explizit eingefügten physikochemischen Information in Verbindung bringen ließ. Dieses Ergebnis kann damit erklärt werden, dass bereits die reine Einteilung in Nukleobasen implizite physikochemische Informationen trägt. Durch die geringe Größe des zugrundeliegenden Alphabetes von lediglich 4 bis 8 ist der Anteil dieser impliziten physikochemischen Information nicht unerheblich. Entsprechend der Ergebnisse ist er hinreichen groß, sodass eine weitere explizite Einbringung physikochemischer Informationen nur noch marginale Effekte gezeigt hat. Durch die Nutzung von Positionsinformation, welche die nachbarschaftlichen Verbindungen der Nukleobasen mit in die Betrachtung einbezieht, stehen dem Regressionsalgorithmus auch die Umgebungsbedingungen der jeweiligen Nukleobasen zur Verfügung. Auf diese Weise ist dieser nicht mehr auf die separierte Betrachtung beschränkt. Allgemein kann also festgestellt werden, dass lokale Nukleobasendeskriptoren nicht explizit mit weiterer physikochemischer Information angereichert werden müssen, da der Verbesserungseffekt bei der Erzeugung von PLS-Regressionsmodellen vernachlässigbar ist. Andererseits wurde gezeigt, dass eine solche, explizite Anreicherung der Deskriptoren keinen negativen Effekt in der Regression zeigt. Es ist jedoch die Frage, ob der nahezu ausbleibende Effekt den hohen Aufwand rechtfertigt, der für die Vorbereitung und Integration der physikochemischen Informationen notwendig ist.

Parametrisierung von n-Gramm-Deskriptoren Die auf n-Grammen basierte Beschreibung nutzt Sequenzfragmente, um relative Positionsinformationen ohne den Zwischenschritt der Nukleobasendeskriptoren zu kodieren. Unter Einbezug aller verfügbarer Deskriptoren konnte mit diesem Verfahren ein Ergebnis erzielt werden, welches vergleichbar mit denen der anderen positionsinformationsbasierten Beschreibungsverfahren ist, wie Abbildung 3.6a zeigt. Weder die Veränderung der Wortgröße noch die Verwendung von SSI bei der Erzeugung der n-Gramme führte dabei zu einer maßgeblichen Verbesserung der Vorhersagegenauigkeit der zugehörigen Regressionsmodelle. Siehe dazu Abbildung 3.5a. Die Durchführung der *Feature Selection* zeigte jedoch sowohl bei der Verwendung von SSI als auch bei Erhöhung der Wortlänge einen deutlichen Effekt. Die erreichte Verbesserung der Modellvorhersage war dabei abhängig von der Anzahl der Deskriptoren der jeweiligen Beschreibungssets, da mit zunehmender Menge an Deskriptoren mehr Raum für die Optimierung bei der Auswahl geschaffen wurde. In diesem Sinne wirkte sich die Vergrößerung der Wortlänge signifikant positiv auf das Ergebnis aus ($p < 0.01$), wobei sich eine Begrenzung bei $n = 4$ angedeutet hat. Weiterhin konnte mit hoher Signifikanz ($p < 0.001$) festgestellt werden, dass die Verwendung von SSI bei allen getesteten Wortlän-

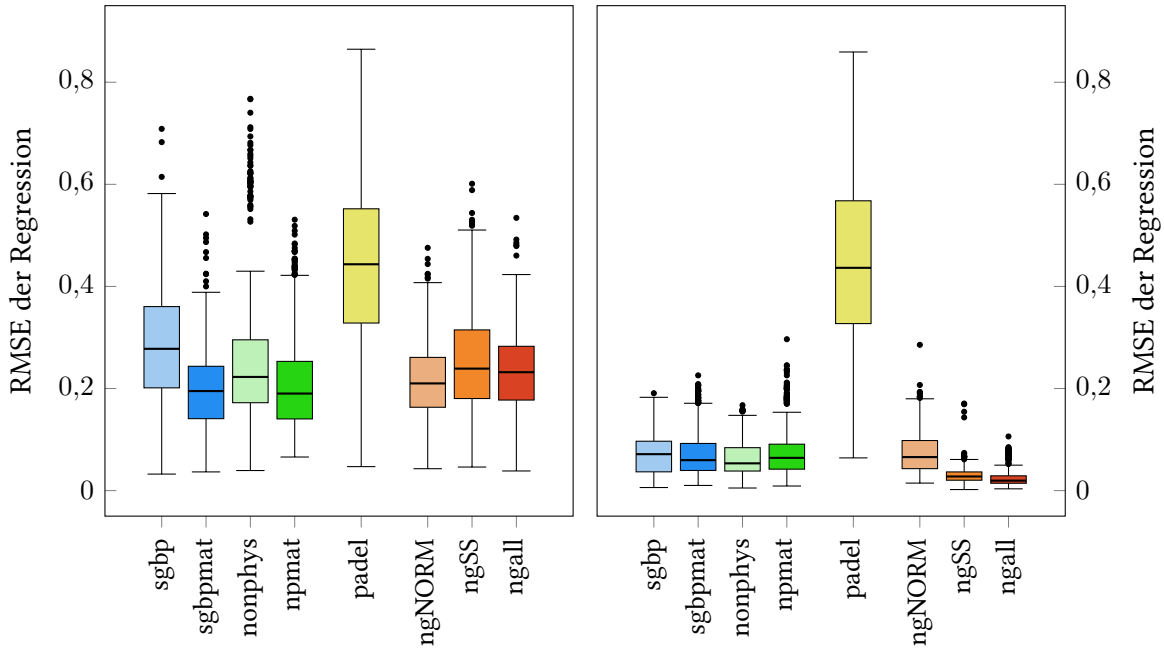


Verwendung aller Deskriptoren **(a)** **(b)** Nach *Feature Selection*

Abb. 3.5: Vergleich zwischen verschiedenen n-Gramm-Deskriptoren. Vergleichsgrößen sind die kreuzvalidierten RMSE-Werte der PLS-Regressionsmodelle, die auf den jeweiligen Beschreibungssets trainiert wurden. Die Farbgebung und Benennung der Beschreibungssets richtet sich nach den Vorgaben in Tabelle 3.4. Der Effekt der unterschiedlichen Wortlängen (Gruppierung) und der Einbringung von SSI (Farbintensität) auf die Vorhersagekraft der Regressionsmodelle ist nach der *Feature Selection* signifikant stärker ausgeprägt als davor. Sowohl der Einsatz längerer n-Gramme als auch die Nutzung von SSI führt zu Regressionsmodellen mit geringeren Vorhersagefehlern.

gen zu einer deutlichen Verringerung des RMSE beitrug. Dies kann durch den Vergleich der Beschreibungssets `ngNORM`n und `ngSS`n in Abbildung 3.5b nachvollzogen werden. Der gemeinsame Einsatz von n-Grammen mit und ohne SSI lieferte hingegen nicht in allen Fällen eine weitere Verbesserung. Besonders hervorzuheben sind die Beschreibungssets mit gemischten Wortlängen ($n = 1 \ldots 4$). Zwar waren die Mittelwerte ihrer Vorhersagefehler mit denen derer vergleichbar, die eine feste Länge von $n = 4$ aufwiesen, die Varianz lag jedoch im gemischten Fall geringer.

Als positionsunabhängige, lokale Sequenzfragmente passen n-Gramme ihrer Natur nach am besten zum Prinzip der Bindemotive der Promotoren. Sie eignen sich daher bei geeigneter Zusammenstellung besser zur Beschreibung der Promotoren als die unterschiedlich transformierten Nukleobasendeskriptoren. Dies steht in Übereinstimmung mit vorausgehenden Studien, welche für Promotorensequenzen erfolgreich binäre Klassifikatoren auf Basis künstlicher neuronaler Netzwerke mit n-Gramm-Deskriptoren trainierten [252; 253]. Da die n-Gramme die lokalen biologischen Strukturen abbilden, die in den darunterliegenden Sequenzen spezifische Bindungsfunktionen erfüllen, hängt deren Aussagekraft in der PLS-Regression stark von der biologischen Relevanz der erfassbaren Fragmente ab. Dementsprechend weisen zu klein gewählte n-Gramme nicht die Unterscheidungskraft auf, die für die Regression nötig ist. Im umgekehrten Fall leidet die Generalisierungsfähigkeit der Modelle unter einer zu hohen Wortlänge, sodass eine geeignete Lösung gefunden werden muss. Neben der Komponente der biologischen Relevanz ist jedoch auch die Menge der erzeugten Deskriptoren algorithmisch relevant, denn sowohl die Erhöhung der Wortlänge als auch die Integration von Sekundärstrukturinformationen führt zu einem massiven Anstieg der erzeugten Deskriptorenzahl. Dieser Anstieg entfällt nicht allein



Verwendung aller Deskriptoren **(a)** **(b)** Nach *Feature Selection*

Abb. 3.6: Überblick über eine repräsentative Auswahl der Beschreibungssets sowohl vor als auch nach der *Feature Selection*. Vergleichsgrößen sind die kreuzvalidierten RMSE-Werte der PLS-Regressionsmodelle, die auf den jeweiligen Beschreibungssets trainiert wurden. Die Farbgebung und Benennung der Beschreibungssets richtet sich nach den Vorgaben in Tabelle 3.4. Neben der großen Gruppe ähnlich gut performender Beschreibungssets stechen zwei weitere Gruppen heraus. Zum einen die PaDEL-Deskriptoren, welche nicht geeignet sind, um die Promotorsequenzen sinnvoll zu beschreiben. Besonders erkennbar ist dies daran, dass die PaDEL-Deskriptoren durch die *Feature Selection* keine Verbesserung erfahren. Zum Anderen hebt sich die Gruppe der mit SSI angereichten n-Gramm-Deskriptoren durch besonders niedrige Vorhersagefehler ab. Die hohe statistische Signifikanz ($p < 0.001$) des besten Sets wird durch ** angegeben.

auf relevante sondern zu einem großen Teil auf irrelevante Deskriptoren. Diese treten jedoch in der algorithmischen Verarbeitung als Störgrößen auf und beeinflussen die Güte der Regression damit zunehmend negativ. Offensichtlich haben sich diese beiden Effekte vor der Anwendung der *Feature Selection* ausgeglichen, sodass die unterschiedlichen n-Gramm-Beschreibungssets Modelle ähnlicher Güte lieferten. Durch die Auswahl relevanter Deskriptoren wurde der negative Effekt weitestgehend eliminiert, sodass die Güte der Modelle entsprechend der Beschreibung des positiven Effekts anstieg, wie in Abbildung 3.5b sichtbar wird.

Abschließende Betrachtung Auch wenn sich beim Training der Regressionsmodelle mit allen in den jeweiligen Beschreibungssets vorhandenen Deskriptoren mit Ausnahme der PaDEL-Deskriptoren kaum Unterschiede zeigten, ließ die *Feature Selection* einen Einblick in die Anwendbarkeit der unterschiedlichen Beschreibungsformen zu. So manifestierten sich unter den Beschreibungssets nach der *Feature Selection* in Abbildung 3.6b drei Gruppen. Die PaDEL-Deskriptoren bilden dabei als Deskriptoren ohne Positionsinformation die Gruppe mit dem höchsten RMSE. Da weder die Einbringung der expliziten physikochemischen Informationen in die Nukleobasendeskriptoren, noch die Art der Positionsinformation maßgeblichen Einfluss zeigen, bilden die Beschreibungssets der indirekten Beschreibung eine weitere Gruppe, welcher zusätzlich die n-Gramm-Deskriptoren ohne Sekundärstrukturinformationen angehören. Die n-Gramme, welche mit Sekundärstrukturinformationen angereichert worden sind, zeigen

einen kleineren RMSE und werden damit als beste Gruppe zusammengefasst. Im Rahmen dieser vergleichenden Evaluation lieferte das n-Gramm-basierte Set `ngall` den geringsten RMSE ($p < 0.001$) und damit die beste Beschreibung der zugrundeliegenden Promotorsequenzen.

Die bisherigen Ergebnisse zeigen, dass eine globale Beschreibung am wenigsten geeignet für die PLS-Regression der Wirkungsstärken von Promotoren ist. Besonders dann, wenn Positionsinformationen gänzlich fehlen oder, wie bei den PaDEL-Deskriptoren der Fall, nur in einer unpassenden Größenordnung vorliegen, kommen sowohl der Regressionsalgorithmus als auch die *Feature Selection* an ihre Grenzen. Der Art und Weise ihrer Funktion geschuldet wird die Bindung der Promotoren von zwei wichtigen Sequenzmotiven vermittelt [212–214], deren korrekte numerische Abbildung ohne das Erhalten einer Form der Positionsinformation nicht möglich ist. Das beste Ergebnis wurde durch die Kombination verschiedener n-Gramm-Beschreibungssets erzielt. Die Zusammenstellung im Beschreibungsset `ngall` umfasst nicht nur verschiedene Wortlängen ($n = 1 \ldots 4$), sondern auch die Kombination der ursprünglichen n-Gramme mit den um SSI angereicherten Varianten. Die große Flexibilität, die durch diese Komposition erreicht wird, erlaubt eine optimale *Feature Selection* und führte daher in den Regressionsmodellen zu einem kleinstmöglichen Vorhersagefehler. Da diese hohe Flexibilität jedoch auch das Risiko einer Überanpassung birgt, wurde das am besten befundene Modell im weiteren auf seine Plausibilität hin überprüft.

3.3.2 Überprüfung der Plausibilität

Plausibilität ist eine grundlegende Anforderung an jedes zuverlässige mathematische Modell. Nachdem PLS-Regressionsmodelle anhand der Beschreibungssets trainiert wurden und eine Rangliste der Anwendbarkeit dieser Sets entstand, folgte die Überprüfung des besten gefundenen Regressionsmodells. Als Referenz für die Plausibilitätsprüfung wurden die bekannten biologischen Charakteristika der Promotoren herangezogen. Da die relative Wirkungsstärke der Promotoren als Zielgröße der Regression verwendet wurde, konnte der bekannte Einfluss der beiden Sequenzmuster RPS und *Pribnow Box* zur Überprüfung der Modellrelevanz herangezogen werden. Sie wirken sowohl auf das Zustandekommen der Promotorbindung als auch auf die Leistungsfähigkeit des Promotors in der Transkription. Auf diesen Sachverhalt bezogen sollte ein plausibles Regressionsmodell sensibel auf die Anwesenheit, Abwesenheit, Modifizierung und Positionierung der beiden Motive reagieren. Diese sollten entsprechend als positive und negative Beiträge der beteiligten Nukleobasen sichtbar werden.

Erste Prüfung des Modells Zu diesem Zweck wurden die Beteiligungen am Regressionsergebnis pro Nukleobase bestimmt und zur einfacheren Übersicht in Abbildung 3.7 als Heatmap aufgetragen. Verglichen mit dem gegebenen Sequenzdatensatz tendierten die nukleobasenweisen Beiträge in der Tat dazu, die Anwesenheit und Abwesenheit der Konsensussequenz `TTGACA` der RPS widerzuspiegeln. Während die -35-Region jedoch deutlich aus der Abbildung hervorgeht, fehlt ein solches graphisches Äquivalent für die *Pribnow Box*, welche sich in der -10-Region befindet, gänzlich. Neben den nukleobasenweisen Beiträgen wurden nun auch die unverrechneten β^*-Faktoren betrachtet, die den Einfluss der n-Gramme wiedergeben. Diese Betrachtung lieferte ein übereinstimmendes Ergebnis. Durch die *Feature Selection* wurden eine Reihe von 3- und 4-Grammen gewählt, die an der Bildung der RPS beteiligt waren. Die β^*-Faktoren dieser n-Gramme zeigten durchgehend positive Beiträge. Im Gegensatz dazu wurden nur sehr wenige der möglichen n-Gramme unter den ausgewählten *Features* gefunden, die zur Bildung der *Pribnow Box* beitrugen. Ihre β^*-Faktoren leisteten bedeutend kleinere positive Beiträge zur Regression.

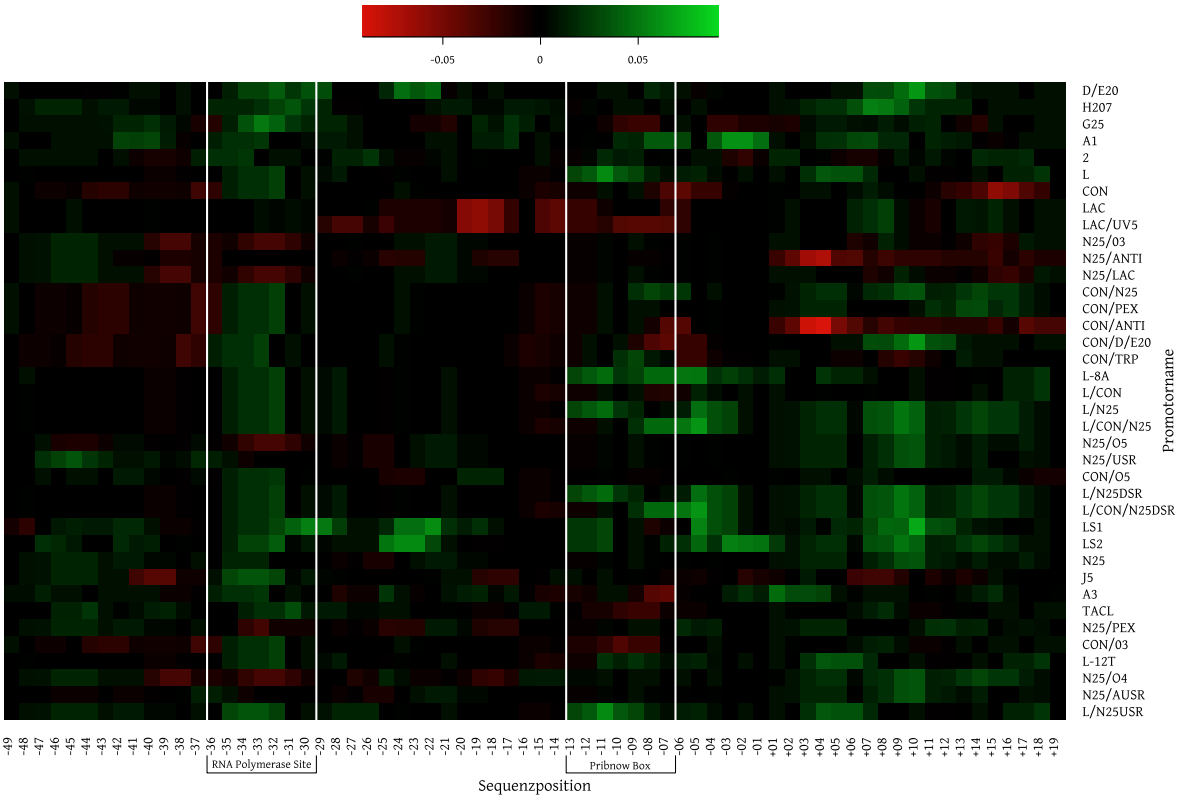


Abb. 3.7: Visualisierung der nukleobasenweisen Beiträge zum Regressionsergebnis. Das PLS-Regressionsmodell wurde dafür unter Anwendung einer *Feature Selection* mithilfe des gemischten n-Gramm-Beschreibungssets `ngall` trainiert, welches sich im Vergleich als bestes herausgestellt hatte. Für jede Nukleobasenposition wird der Beitrag zum Regressionsergebnis in der Heatmap durch einen Farbwert dargestellt. Der Farbbereich reicht dabei von grün (positiver Beitrag) über schwarz (kein Beitrag) bis hin zu rot (negativer Beitrag). Neben einer schwachen, aber stimmigen Ausprägung der RPS ist die *Pribnow Box* in der Darstellung nicht repräsentiert.

Erweiterung der Beschreibung Da eine bessere Repräsentation der *Pribnow Box* zu einem natürlicheren Verständnis des Modellierungsergebnisses beigetragen hätte, wurden einige Versuche unternommen, durch Modifikation des Deskriptorensets eine bessere Abbildung der *Pribnow Box* in den Regressionsmodellen zu erwirken. Der Grundgedanke dabei war, dass eine Erweiterung um zusätzliche Informationen das Regressionsmodell befähigen würden, das zweite Bindungsmotiv und damit die biologische Realität erfolgreich abzubilden. Im ersten Anlauf wurde die maximale Wortlänge auf $n = 6$ vergrößert, was der Länge der beiden betrachteten Motive entspricht. Das resultierende Beschreibungsset, welches nun n-Gramme der Wortlängen $n = 1..6$ enthielt, wurde schließlich zum Trainieren eines Regressionsmodells genutzt. Leider führte diese Erweiterung weder zu einer deutlichen Ausprägung der *Pribnow Box* in der grafischen Darstellung, noch zu einer besseren Repräsentation in den ausgewählten n-Grammen der *Feature Selection*. Ferner konnte auch keine Verbesserung bei der erzielbaren Regressionsgenauigkeit festgestellt werden. Auch die ausschließliche Nutzung der 6-Gramme führte nicht zum Erfolgt. Im zweiten Anlauf wurden die n-Gramme durch eine kleine Menge an absoluter Positionsinformation ergänzt, um deren Einfluss zu überprüfen. Um dies zu erreichen, wurden die Sequenzen vor der Bestimmung der n-Gramm-Häufigkeiten in wenige Kompartimente einer festgelegten Größe unterteilt. Ein n-Gramm wurde in der weiteren Betrachtung nicht mehr nur durch seine Nukleotidsequenz beschrieben, sondern zusätzlich noch durch die Nummer des zugehörigen Kompartiments. Die in dieser Annotation in die n-Gramme integrierte absolute Po-

sitionsinformation war damit unabhängig von der Sequenzlänge. Das gewünschte Ergebnis, die bessere Herausbildung der *Pribnow Box* während der Modellierung, konnte diese Erweiterung jedoch ebenfalls nicht bewirken.

Erweiterung des Datensatzes Da der originale Datensatz nur Sequenzen von funktionsfähigen, jedoch keine von funktionsunfähigen Promotoren enthielt, war der verfügbare Bereich von Wirkungsstärken, der den Regressionsmodellen während des Trainings zur Verfügung stand, sehr eingeschränkt. Dies lässt die Vermutung aufkommen, dass die schlechte Manifestation der *Pribnow Box* in den β^*-Faktoren und der grafischen Darstellung der nukleobasenweisen Beiträge darauf hindeutet, dass diese stärker an der Unterscheidung zwischen positiven und negativen Sequenzen beteiligt ist. Ausgehend von dieser Annahme wäre nicht nur das erhaltene Ergebnis plausibel gewesen, sondern zusätzlich die bessere Herausbildung der *Pribnow Box* durch die Einführung von funktionsunfähigen Sequenzen in den Datensatz überprüfbar. Zu diesem Zweck wurden 33 zufällige DNA-Sequenzen erzeugt und in den originalen Sequenzdatensatz eingefügt. Da die Wahrscheinlichkeit, zufällig bindende Sequenzen zu erzeugen, hinreichend gering war, konnten die sequenzspezifischen Werte für die Wirkungsstärke auf einen festen, symbolischen Wert von -5 gesetzt werden. Dieser drückte in Ermangelung einer konkreten Quantifizierung aus, dass die Sequenzen keine wirksamen Promotoren waren. Nach der Erzeugung eines Regressionsmodells wurden erneut die nukleobasenweisen Beiträge zum Regressionsergebnis bestimmt und, wie in Abbildung 3.8 ersichtlich, grafisch aufgetragen. Das Vorhandensein, die Form und auch die genaue Positionierung der beiden Sequenzmotive, *Pribnow Box* und RPS, wurden in dieser Darstellung nun korrekt repräsentiert. Zum Vergleich kann Tabelle 3.3 herangezogen werden.

Bewertung der Plausibilität Da der originale Datensatz also nur Sequenzen funktionsfähiger Promotoren enthielt, fand während der Regression eine Anpassung an einen sehr eingeschränkten Bereich der Wirkungsstärken statt. Das resultierende Modell war folglich darauf spezialisiert, interne Variationen der Sequenz zu unterscheiden, die für die Variationen der Wirkungsstärke im Datensatz verantwortlich waren. Eine Unterscheidung zwischen funktionsfähigen und funktionsunfähigen Promotoren war mit diesem Modell hingegen nicht möglich. Die Beobachtung, dass erst nach Hinzufügen von expliziten Negativbeispielen beide Bindemotive korrekt repräsentiert wurden, unterstützt die Vermutung, dass die *Pribnow Box* essentiell für die Promotoraktivität ist und in der Initiation der Transkription eine nicht ersetzbare Rolle einnimmt. Die zur *Pribnow Box* gehörenden Sequenzfragmente wurden nicht durch die *Feature Selection* ausgewählt, weil sie im betrachteten Kontext ausschließlich funktionsfähiger Promotorsequenzen kaum Unterscheidungskraft besaßen. Im Gegenzug kann dem Ergebnis entnommen werden, dass die RPS stärker für Variationen der Wirkungsstärken innerhalb funktionsfähiger Promotoren verantwortlich ist. Diese Schlussfolgerung wird durch eine Reihe existierender Studien unterstützt. In diesen wurden unter anderem gezeigt, dass die RPS zwar an der Funktion der Promotoren beteiligt ist, sie wurde jedoch nicht als essentiell notwendig für die Einleitung der Transkription befunden. Da das Ersetzen des Hexamers im -35-Bereich die Zeit zur Bildung des offenen Komplexes deutlich erhöht, kann davon ausgegangen werden, dass dieses Bindemotiv einen großen Einfluss auf die Effektivität der laufenden Transkription ausübt. Seine Abwesenheit kann jedoch unter bestimmten Umständen durch eine spezifische Erweiterung der *Pribnow Box* ausgeglichen werden [216; 254–256]. Berücksichtigt man also diese Betrachtung der biologischen Hintergründe, so hat das erste Modell korrekterweise die RPS hervorgehoben,

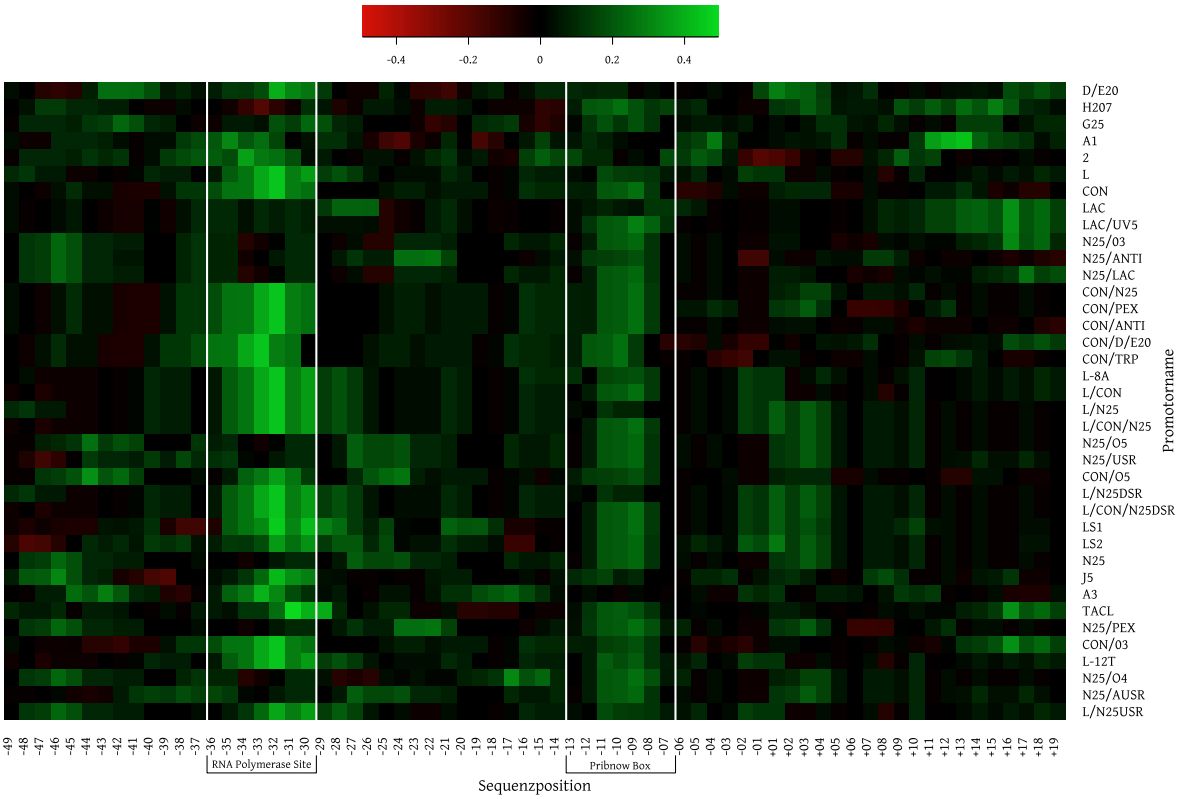


Abb. 3.8: Visualisierung der nukleobasenweisen Beiträge zum Regressionsergebnis. Das PLS-Regressionsmodell wurde dafür unter Anwendung einer *Feature Selection* mithilfe des gemischten n-Gramm-Deskriptorensets `ngall` auf einem erweiterten Sequenzdatensatz trainiert, welcher zusätzlich 33 zufällige DNA-Sequenzen als Repräsentanten funktionsunfähiger Promotoren enthielt. Für jede Nukleobasenposition wird der Beitrag zum Regressionsergebnis in der Heatmap durch einen Farbwert dargestellt. Der Farbbereich reicht dabei von grün (positiver Beitrag) über schwarz (kein Beitrag) bis hin zu rot (negativer Beitrag). Die eingefügten, funktionsunfähigen Sequenzen sind aus Gründen der Übersicht nicht in die Heatmap aufgenommen worden. Die zwei wichtigen Bindemotive der Promotoren bilden sich nun deutlich und korrekt in der Darstellung heraus.

da diese für die Variationen der Wirkungsstärke im Datensatz verantwortlich war. Nach dem Einfügen von Negativsequenzen wurden beide Motive entsprechend repräsentiert, da nun auch die Unterscheidung von funktionsunfähigen Sequenzen notwendig war. Die beiden betrachteten Modelle werden also als plausibel befunden.

Weitere Beiträge Die Modelle zeigen aber auch, dass nicht nur die beiden Bindemotive, sondern auch Sequenzfragmente dazwischen und außerhalb Beiträge zur Regression geleistet haben. Das wird in den Abbildungen 3.7 und 3.8 deutlich. Tatsächlich kann dies zum Teil das Ergebnis der positionsunabhängigen Natur der n-Gramme sein. Ohne die absolute Positionsinformation können n-Gramme, die Teile der Binderegion eines Motivs abbilden, beim späteren Abbilden nicht mehr exakt diesem zugeordnet werden. Kommen gleiche Fragmente an anderer Stelle der Sequenz vor, so haben diese in der späteren Analyse automatisch Anteil an den ermittelten Beiträgen. Da es unwahrscheinlich ist, dass alle Beiträge, die außerhalb der Motive liegen, auf diese Weise zustande gekommen sind, kann angenommen werden, dass die Interaktion zwischen Promotor und RNA-Polymerase nicht allein, jedoch hauptsächlich von den zwei Bindemotiven vermittelt wird. Diese Funktion der *Spacer*-Regionen in der Transkription wurde von Singh *et al.* am Beispiel von bakteriellen Promotoren untersucht und bestätigt [257]. Beachtet man also

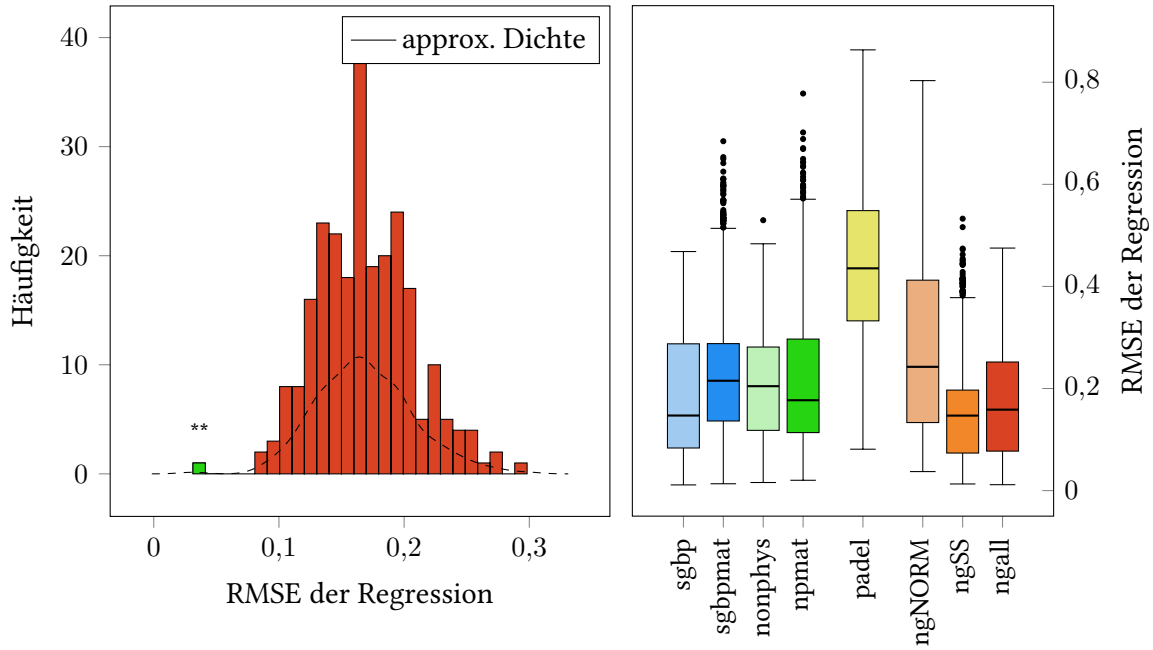
die möglichen Einflüsse der *Spacer*-Regionen, die Kreuzeffekte der n-Gramme und die geringe Größe des Datensatzes, so sind auch die Modellbeiträge außerhalb der eigentlichen Bindemotive gerechtfertigt.

3.3.3 Verhältnismäßigkeit

Gerade in der Kombination von PLS und *Feature Selection* liegt jedoch eine große Gefahr. Bereits die Fähigkeit des Verfahrens PLS, Eingabedaten zu verarbeiten, deren Anzahl unabhängiger Variablen die Anzahl der vorhandenen Trainingsdatensätze weit übersteigt, birgt das Risiko einer Überanpassung, dem sogenannten *Overfitting*. Ist ein Vorhersagemodell derart überangepasst an die zugrundeliegenden Eingabedaten, so führt dies zu einer mangelnden bis fehlenden Generalisierungsfähigkeit. Im ungünstigsten Fall lernt der Trainingsalgorithmus auf zufälligen oder entstellten Eingabedaten mit der gleichen Präzision wie auf den echten Eingabedaten, sodass die erhaltenen Regressionsmodelle keine Aussagekraft besitzen [234; 258]. Dieses Risiko erfährt eine deutliche Minderung durch die eingesetzte KV, welche durch wechselnde, randomisierte Teilmengen bei Training und Validierung eine gewisse Generalisierung der hervorgebrachten Regressionsmodelle erfordert [258; 259]. Analysen haben jedoch gezeigt, dass die gewöhnliche k-fache KV ähnlich wie die *Leave-One-Out* (LOO)-KV ohne die Einhaltung bestimmter Nebenbedingungen nicht geeignet sind, um *Overfitting* zu vermeiden. In einer Studie[260] wurde in diesem Zusammenhang das Verhalten der LOO-KV analysiert, welches nicht zu einer statistisch konsistenten Wahl des wahren Modells führte. Selbiges gilt auch für die k-fache KV. Um diese Probleme zu beheben, wird eine sogenannte *Leave-Multiple-Out* (LMO)-KV empfohlen, bei der die Grundmenge der Daten mehrfach zufällig in Trainings- und Validierungsdaten unterteilt wird, wobei ein größerer Teil der Daten jeweils zur Validierung verwendet werden soll [261].

Beurteilung der eingesetzten Verfahren Die in dieser Arbeit angewendete zehnfache KV erfüllt damit an sich nicht die Anforderung, eine anteilsmäßig größere Validierungsdatenmenge zu nutzen, weil dies bei der geringen Gesamtdatenlage zu einer unzureichenden Menge an Trainingsdaten führen würde. Durch die 50 randomisierten Wiederholungen dieses Verfahrens wird jedoch eine hohe Abdeckung der Teilmengen erreicht, wie sie auch beim LMO-KV angestrebt wird. Einer Überanpassung auf eine konkrete Zehner-Teilung der Datenmenge wird daher vorgebeugt. Da die *Feature Selection* diesen Validierungsmechanismus jedoch nur im internen Prozessteil verwendet, besteht weiterhin die annehmbare Möglichkeit einer Überanpassung an die vorhandenen Eingabedaten [262]. So wurde unter anderem in einem Versuch festgestellt, dass eine genetische *Feature Selection* auch einige zufällig generierte Variablen aus einem Datensatz auswählte [261; 263]. Es besteht also dennoch die Notwendigkeit, die erhaltenen Regressionsmodelle auf *Overfitting* hin zu überprüfen. Eine bewährte Methode ist das sogenannte Y-*Scrambling*, ein Permutationstest, bei dem das Verhalten des Regressionsverfahrens unter Permutation der Zielgröße (Y) analysiert wird. Ergeben sich in einer solchen Prüfung trotz unnatürlicher Permutation der Zielgröße und damit fehlendem kontextuellem Zusammenhang Modelle, die eine gute oder sogar bessere[264] Vorhersageleistung versprechen, so belegt dies ein starkes *Overfitting* der eingesetzten *Feature Selection*-Strategie [261; 265; 266].

Durchführung des Permutationstests Für das beste Deskriptorenset `ngall` wurde ein solcher Permutationstest durchgeführt. Dazu wurden aus dem ursprünglichen Datensatz unter Beibehaltung der Deskriptorenwerte und zufälliger Permutation der Zielgröße logRS 250 modifizierte Datensätze zur Gegenprobe erstellt. Für jeden dieser Datensätze wurde der komplette

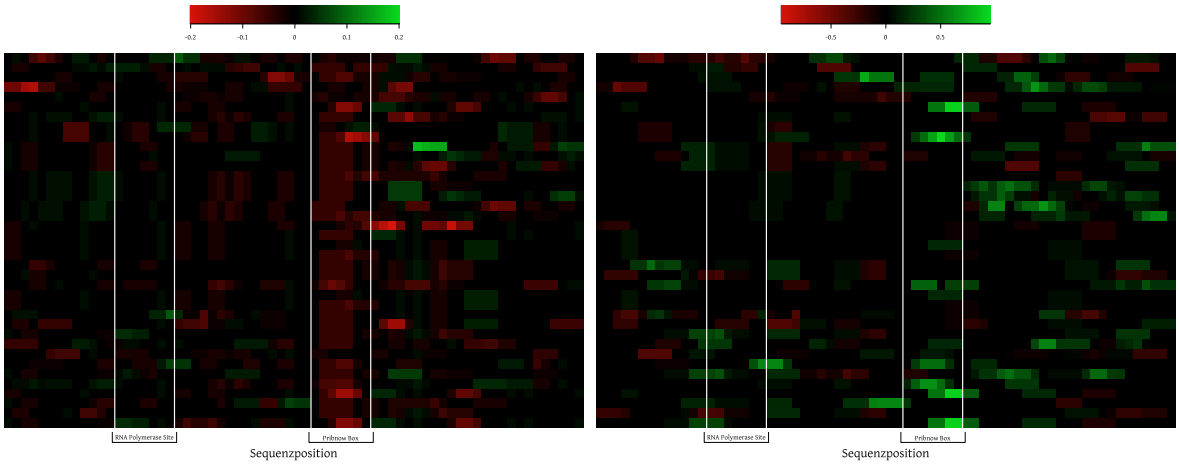


Verteilung der Modellfehler **(a)** **(b)** Auswahl der Beschreibungssets

Abb. 3.9: (a) Verteilung der Modellfehler nach Erzeugung von 250 Modellen auf Basis verschiedener Randomisierungsläufe (rot) im Vergleich mit Originalmodell (grün), jeweils mit dem Deskriptorenset `ngall`. Es wird deutlich, dass die Modelle, welche nach Permutation der Zielgröße erhalten wurden, deutliche größere Modellfehler aufweisen (** p<0.01). (b) Überblick über eine repräsentative Auswahl der Beschreibungssets nach der *Feature Selection* nach zufälliger Permutation der Zielgröße. Auch wenn die tendenziellen Wertungen erhalten bleiben, führt die Randomisierung der Zielgröße zu einer deutlichen Vergrößerung der Modellfehler.

Optimierungszyklus mit Regression, Kreuzvalidierung und *Feature Selection* durchlaufen. Die erhaltene Verteilung der durchschnittlichen Modellfehler ist in Abbildung 3.9a dargestellt. Es wird deutlich, dass das Regressionsmodell, welches auf dem originalen Datensatz trainiert wurde, einen wesentlich kleineren mittleren Modellfehler aufweist. Das gleiche Ergebnis trifft beim Vergleich auch auf die Quartile, den Median und die Verteilung von Ausreißern zu. Der bereits optisch zu erkennende Unterschied zwischen dem Modellfehler des Originalmodells mit $0{,}031$ und der approximierten Verteilungsfunktion der Modellfehler bei zufälliger Zuordnung der Zielgröße logRS um $0{,}169$ ließ sich statistisch belegen. So ergab ein einseitiger Zweistichproben-t-Test eine sehr hohe Signifikanz ($p < 0.001$) des Unterschieds. Es kann daher gefolgert werden, dass die *Feature Selection* ordnungsgemäß funktioniert hat, ohne eine Überanpassung hervorzurufen.

Detailbetrachtung des Permutationseffektes Weiterhin wurde für alle am ursprünglichen Vergleich beteiligten Deskriptorensets jeweils ein modifizierter Datensatz nach dem gleichem Prinzip erstellt. Dieser durchlief ebenfalls den Optimierungszyklus und diente nach der Betrachtung der quantitativen Effekte zum exemplarischen Vergleich der qualitativen Unterschiede. Aus Abbildung 3.9b kann entnommen werden, dass zwar die Grundtendenz der Eignung der unterschiedlichen Deskriptorensets (siehe dazu im Vergleich Abbildung 3.6b) erhalten bleibt, die jeweiligen Modelle jedoch wesentlich größere Fehler aufweisen. Besonders deutlich werden die Unterschiede, wenn aus den Regressionsmodellen wie in Abbildungen 3.7 und 3.8 bereits am nicht-permutierten Modell geschehen auf die Beteiligung der einzelnen Nukleobasen an der Modellbildung geschlossen wird. Während die Plausibilität des nicht-permutierten Modells



Modell ohne Negativproben (vgl. Abb. 3.7). **(a)** **(b)** Modell mit Negativproben (vgl. Abb. 3.8).

Abb. 3.10: Visualisierung der nukleobasenweisen Beiträge zum Regressionsergebnis nach Randomisierung der Zielgröße. Das PLS-Regressionsmodell wurde dafür unter Anwendung einer *Feature Selection* mithilfe des gemischten n-Gramm-Deskriptorensets `ngall` trainiert. Für jede Nukleobasenposition wird der Beitrag zum Regressionsergebnis in der Heatmap durch einen Farbwert dargestellt. Der Farbbereich reicht dabei von grün (positiver Beitrag) über schwarz (kein Beitrag) bis hin zu rot (negativer Beitrag). Nach Randomisierung bilden sich die zwei wichtigen Bindemotive der Promotoren nicht in der Darstellung heraus.

auf diese Weise biologisch hinterlegt werden konnte, zeigt das Modell nach Permutation der Zielgröße logRS wie in Abbildung 3.10 zu erkennen keinerlei relevante Korrelation zum Bindungsmodus der Promotoren. Die Bestätigung der Plausibilität des nicht-permutierten Modells war damit ebenfalls kein Ergebnis des Zufalls.

3.3.4 Abschließende Betrachtung

Nach deren Vorstellung wurde in diesem Kapitel der evaluierende Vergleich der unterschiedlichen Deskriptorensets an einem Datensatz von Promotorsequenzen durchgeführt. Dieser zeigte, dass die Güte der Beschreibung dabei zum größten Teil aus der Kombination von impliziter physikochemischer Information und Positionsinformation hervorgeht. Die Einteilung in Nukleobasen scheint dabei hinreichend viel implizite physikochemische Information zu enthalten, sodass der Effekt der zusätzlichen Einbringung vernachlässigbar ist. Besonders wichtig hat sich im Vergleich die Möglichkeit der Vernetzung dieser physikochemischen Information herausgestellt, die durch die Verwendung von Positionsinformationen ermöglicht wird. Fehlte diese gänzlich oder lag sie in einer unpassenden Größenordnung vor, so zeigte sich trotz *Feature Selection* nur eine schlechte Anpassung der Regressionsmodelle an die Eingabedaten.

Für die Beschreibung von Nukleinsäuren sollten daher stets Deskriptoren verwendet werden, die Positionsinformationen in der kompatiblen Größenordnung tragen. Von der expliziten Einbringung physikochemischer Informationen in Nukleobasendeskriptoren wird zwar nicht prinzipiell abgeraten, jedoch sollte der Aufwand dieses Schrittes ins Verhältnis zum erwarteten Nutzen gebracht werden. Globale, physikochemische Deskriptoren, welche primär zur Beschreibung von kleineren Molekülen entwickelt wurden, weisen bei Nukleinsäuren hingegen keine hinreichende Beschreibungsgüte auf. Besonders, wenn mit lokalen Bindemotiven auf den Nukleinsäuren gerechnet wird, zeigen sich die Stärken einer Beschreibung mit n-Grammen. Die Veränderung der eingesetzten Wortlängen, die Einbringung von Sekundärstrukturinformationen und die Kombination dieser Modifikationen erwiesen sich als wirksame Mittel zur

Steigerung der Beschreibungsgüte. Im Rahmen dieser Untersuchung zeigte sich ein zusammengestelltes Beschreibungsset aus n-Gramm-Deskriptoren als ideale Form der Beschreibung der Promotorsequenzen, welche die biologischen Charakteristika der Promotoren korrekt abbildete. Die erlangten Erkenntnisse können auch auf andere Formen der funktionellen Nukleinsäuren übertragen werden, da sie nach dem übereinstimmenden Prinzip der molekularen Erkennung spezifischer Subsequenzen oder Substrukturen agieren. Die Erkenntnisse dieser Evaluation können daher im Speziellen auch auf die Aptamere angewendet werden.

4 Mustersuche in biologischen Sequenzen

Bereits in Kapitel 3 wurden n-Gramme als wichtiges Beschreibungselement für Nukleinsäuresequenzen herausgestellt, da diese die auftretenden Bindecharakteristika sinnvoll repräsentieren. Als positionsunabhängige Teilsequenzen definierter Länge bilden sie die Grundform des trivialen Sequenzmusters. Gleichzeitig unterliegen sie jedoch strengen Restriktionen, die ihre Anwendung in der Domäne der molekularen Erkennung einschränken und die Interpretation ihrer Bedeutung erschweren. Die Variabilität in den betrachteten Sequenzabschnitten, die in der molekularen Erkennung besonders im makromolekularen Bereich auftritt, bildet den Kernpunkt dieser Problematik. n-Gramm-basierte Beschreibungsansätze können eine solche Variabilität nur auf indirektem Wege und unter erheblichem Mehraufwand erreichen. In diesem Kapitel wird der allgemeine Begriff eines Sequenzmusters für biologische Sequenzen, speziell für Nukleinsäuresequenzen, genauer definiert. Entsprechend dieser Definition erfolgt die Entwicklung eines Mustersuchverfahrens, welches derartige Muster in großen Sequenzdatensätzen effektiv identifizieren kann.

4.1 Sequenzmuster

In diesem Abschnitt werden die Grundlagen für die weitere Arbeit mit Sequenzmustern gelegt. Dazu wird nicht nur der Musterbegriff sukzessiv definiert, sondern auch die Thematik der Bewertung von Mustern und Musterfunden beleuchtet. Einige Gedanken zur Visualisierung der Muster erhöhen die Zugänglichkeit zu den späteren Mustersuchergebnissen und schließen damit den Themenbereich ab.

4.1.1 Definition der Sequenzmuster

Kernpunkt der Definition komplexer Muster ist die Einführung der Variabilität, mit deren Hilfe die tatsächliche Bindecharakteristik in ihrer Toleranz und intrinsischen Mehrdeutigkeit abgebildet werden kann, ohne die Nachteile bei der Nutzung von Alignmentverfahren und sehr kurzen n-Grammen inkaufzunehmen. Aufbauend auf dem Prinzip der trivialen Sequenzmuster soll eine komplexe Erweiterungsklasse der Sequenzmuster definiert werden. Die Einführung der Variabilität basiert auf dem Prinzip der *Position-Specific Scoring Matrix* (PSSM), einer Matrixdarstellung, welche die Auftretenshäufigkeiten der erlaubten Sequenzbausteine für jede der Musterpositionen erfasst. Über dieses System werden auch auftauchende Lücken als ein Spezialfall eingeschlossen.

Triviale Sequenzmuster Gegeben eines Alphabets $A = \{a_1, \ldots, a_k\}$ der Länge k kann ein triviales Muster $M = (m_1, \ldots, m_n)$ als eine Folge von n Musterpositionen $m_i \in A$ definiert werden. Eine Variabilität im Sinne der partiellen oder vollständigen Austauschbarkeit einzelner Musterpositionen ist über triviale Sequenzmuster nicht modellierbar, in der Praxis jedoch häufig erforderlich. Der Einsatz sehr kurzer trivialer Muster kann in Verbindung mit Alignmentverfahren zwar einen Ansatz zur Modellierung der Variabilität darstellen, ihre Aussagekraft nimmt

jedoch mit der Länge rasch ab. Ausgegangen von einer Gleichverteilung der Nukleobasen in einer zufälligen Sequenz der Länge s ist mit einer Einzelwahrscheinlichkeit einer sequenziellen Übereinstimmung des Musters von k^{-m} zu rechnen. Der daraus abgeleitete Erwartungswert an Musterfunden $E(M)$ (siehe Formel 4.1) sinkt dabei exponentiell mit der Länge des betrachteten Musters. Vereinfacht als Bernoulli-Kette betrachtet ergäbe sich dann die Wahrscheinlichkeit $P(M)$ für mindestens einmaliges Auftreten des Musters in der zufälligen Sequenz, wie in Formel 4.2 gegeben. Aus dieser Näherung kann abgeleitet werden, dass kurze triviale Muster in zufälligen Sequenzen mit relativ hoher Wahrscheinlichkeit unspezifisch anzutreffen sind und daher eine große Neigung zu numerischem Rauschen bergen.

$$E(M) = \frac{s - m + 1}{k^m} \tag{4.1}$$

$$P(M) = 1 - (1 - k^m)^{s-m+1} \tag{4.2}$$

Komplexe Sequenzmuster Der Übergang von trivialen zu komplexen Sequenzmustern erfolgt über die Definition variabler Musterpositionen $m_i^* \subseteq A$, welche jeweils von mehreren Elementen des Alphabets erfüllt werden können. Ein komplexes Sequenzmuster $M^* = (m_1^*, \ldots, m_n^*)$ besteht ausschließlich aus diesen variablen Musterpositionen. Entsprechend der Definition ist die Menge der komplexen jedoch eine echte Obermenge der trivialen Sequenzmuster. Sie modelliert Lücken durch vollständig undefinierte Sequenzpositionen $m_\circ^* = A$. Ein konkreter Fund des komplexen Musters M^* in einem Sequenzdatensatz kann über die relativen Häufigkeitsverteilungen h_{m^*} an seinen Musterpositionen genau beschrieben werden. Als Darstellungsform eignet sich dabei besonders die PSSM [267], deren Bildungsvorschrift in Formel 4.3 gegeben ist. Von dieser können im Weiteren Bewertungsmaße des konkreten Musterfundes abgeleitet werden. Für die Suche eines Musters in einem Sequenzdatensatz eignet sich diese Form jedoch nicht, da mit den angegebenen Häufigkeiten eine sehr hohe Spezifität einher geht. Für die Suche von komplexen Sequenzmustern wird in dieser Arbeit daher die modifizierte Form *Position-Specific Presense Matrix* (PSPM) genutzt. Sie erfasst lediglich das Vorhandensein der einzelnen Elemente des Alphabets an den jeweiligen Musterpositionen (siehe Formel 4.4) und eignet sich daher für die Mustersuche. Theoretisch ließen sich in diese Definition auch Wiederholungen variabler Länge einbringen. Diese sollen jedoch aufgrund der geringen Relevanz und hohen Komplexität in Definition und späterer Implementierung in dieser Arbeit nicht weiter betrachtet werden.

$$\text{PSSM}(M^*) = \begin{pmatrix} h_{m_1^*}(a_1) & \ldots & h_{m_n^*}(a_1) \\ \vdots & \ddots & \vdots \\ h_{m_1^*}(a_k) & \ldots & h_{m_n^*}(a_k) \end{pmatrix} \tag{4.3}$$

$$\text{PSPM}(M^*) = \begin{pmatrix} v_{m_1^*}(a_1) & \ldots & v_{m_n^*}(a_1) \\ \vdots & \ddots & \vdots \\ v_{m_1^*}(a_k) & \ldots & v_{m_n^*}(a_k) \end{pmatrix} \qquad v_{m^*}(a) = \begin{cases} 1 & \forall a \in m^* \\ 0 & \forall a \notin m^* \end{cases} \tag{4.4}$$

4.1.2 Bewertung konkreter Musterfunde

Aus einem Sequenzdatensatz lassen sich nach obiger Definition zahlreiche Muster ableiten. In der algorithmischen Verarbeitung ist es daher besonders wichtig, Musterfunde bewerten zu können, denn nur mithilfe geeigneter Bewertungsmaße können die Rohergebnisse einer

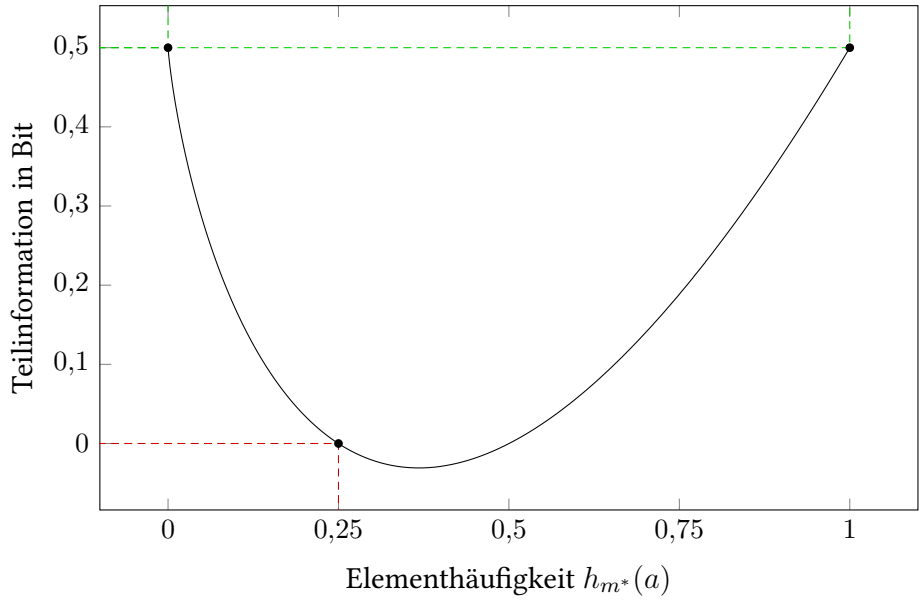


Abb. 4.1: Gezeigt sind die informationellen Teilbeträge einer Musterposition aus Formel 4.6 basierend auf den beteiligten Elementhäufigkeiten beispielhaft für ein Alphabet der Länge 4. An einer trivialen Position werden über die Einzelhäufigkeiten $(\,0\;0\;0\;1\,)$ die maximalen Beträge erreicht (grüne Markierung), während bei einer Gleichverteilung über die Einzelhäufigkeiten $(\,\frac{1}{4}\;\frac{1}{4}\;\frac{1}{4}\;\frac{1}{4}\,)$ keine Information eingeht (rote Markierung).

Mustersuche durch Filterung und Sortierung in eine sinnvolle und nutzbare Ordnung gebracht werden. Im folgenden werden einige mögliche Bewertungskonzepte vorgestellt, die sich jeweils aus der PSSM eines konkreten Musterfundes ableiten lassen.

Informationsgehalt Die Bewertung des Informationsgehaltes leitet sich aus dem Gebiet der Informationstheorie ab. Über die Informationsgehalte $-\log_\psi(p_i)$ aller einzelnen Zeichen mit den Auftrittswahrscheinlichkeiten p_i kann für eine Musterposition die Shannon-Entropie entsprechend Formel 4.5 berechnet werden. Die Wahl des Parameters ψ hat dabei keinen Einfluss auf die Bewertung, sodass dieser in Anlehnung an die binäre Verarbeitungsweise moderner Rechnersysteme mit dem Wert 2 belegt wird. Ungeachtet des Namens handelt es sich bei dieser gewichteten Summe aus einzelnen Informationsgehalten um ein Maß der Ungewissheit der zugrunde liegenden Wahrscheinlichkeitsverteilung [268]. Um daraus ein Maß der Information $I(m^*)$ im eigentlichen Sinne zu erhalten, wird die Differenz zur maximal erreichbaren Entropie des Alphabets $\log_2 k$ entsprechend Formel 4.6 gebildet. Das Ergebnis der Bewertung liegt dabei im Intervall $[0, \log_2 k]$. Wie aus Abbildung 4.1 hervor geht, sind die informationellen Teilbeträge von gleichverteilten Positionen minimal, wohingegen triviale Muster durch die hohen Teilbeträge in der maximalen Bewertung des Informationsgehalts resultieren. Enthält eine Musterposition eine sehr niedrige Information, so könnte diese unter Beachtung eines geeigneten Schwellwertes als Lücke interpretiert werden. Durch Bildung von Summe oder Durchschnitt der positionsweisen Einzelwerte kann zudem auf den Informationsgehalt eines ganzes Musters geschlossen werden, wobei eine Sonderbehandlung von Lücken sinnvoll ist.

$$H_1(m^*) = -\sum_{a \in m*} h_{m^*}(a) \cdot \log_2\left(h_{m^*}(a)\right) \tag{4.5}$$

$$I(m^*) = \log_2 k + \sum_{a \in m^*} h_{m^*}(a) \cdot \log_2\left(h_{m^*}(a)\right) \tag{4.6}$$

Konservierungsgrad Zusätzlich wurde ein Maß zur Beschreibung des Konservierungsgrades einer Musterposition entwickelt. Basis dafür war die Annahme, dass an einer konservierten Musterposition ein Element des Alphabets bezogen auf seine Auftretenshäufigkeit besonders von den anderen abgesetzt ist. Für die Bewertung dieser Häufigkeitsdifferenz wird aus der Musterposition die absteigend sortierte Folge $\widetilde{h}^{m^*}$ der Häufigkeiten ihrer Elemente generiert. Auf dieser Basis wird nicht nur die Differenz des häufigsten Elements zu seinem Nachfolger sondern alle Differenzen bestimmt. Mithilfe einer stark abfallenden Gewichtung wird erreicht, dass die größte Differenz das Bewertungsmaß dominiert, jedoch trotzdem ein geringer Einfluss der weiteren Differenzen einfließt. Auf diese Weise kann einer Teilkonservierung mehrerer Elemente Rechnung getragen werden. Der Konservierungsgrad $K(m^*)$ der Musterposition ergibt sich demnach entsprechend Formel 4.7. Das Ergebnis der Bewertung liegt dabei im Intervall $[0{,}1]$, wobei große Werte eine hohe Konservierung repräsentieren. Zur Beschreibung der Konservierung eines Musters können sowohl der Mittelwert als auch die Summe der Konservierungsgrade der einzelnen Musterpositionen dienen.

$$K(m^*) = \sum_{i=1}^{|m^*|-1} \left(\frac{\sqrt[i]{100 \cdot \left(\widetilde{h}_i^{m^*} - \widetilde{h}_{i-1}^{m^*} \right)}}{100} \right) \tag{4.7}$$

Komplexität Die Bewertung der Komplexität einzelner Musterpositionen und ganzer Muster wurde aus dem Verfahren SEG übertragen, das sich bei der Bewertung einer beliebigen Sequenz maßgeblich auf den sogenannten Komplexitäts-Status-Vektor (KSV) $Z = (z_1, \ldots, z_k)$ bezieht. Für ein schrittweise über die Sequenz bewegtes Fenster erfasst dieser die sortierte Häufigkeitsliste der enthaltenen Zeichen. Zur eigentlichen Bewertung wird neben dem KSV selbst auch die Anzahl möglicher Permutationen des Vektors mit Wiederholung einbezogen. Die Komplexität $X(z)$ ergibt sich schließlich wie in Formel 4.8 gezeigt durch Logarithmierung und Normierung im Intervall $[0{,}1]$, wobei große Werte einer großen Komplexität des Fensterinhalts entsprechen [269; 270]. Das Verfahren wurde zur Bewertung einzelner Musterpositionen modifiziert. Als absteigend sortierte Folge der Auftretenshäufigkeiten aller in der Sequenz vorkommenden Zeichen bezogen auf eine Musterposition entspricht der KSV der bereits eingeführten Größe $\widetilde{h}^{m^*}$. Durch den Einsatz der relativen Häufigkeiten sind die KSVen jedoch reellwertig und damit nicht mit der auf der Menge der natürlichen Zahlen definierten Fakultätsfunktion kompatibel. Zu Behebung dieses Problems wurde die Eulersche Gamma-Funktion $\Gamma(x) \cong (x-1)!\ \forall x \in \mathbb{N}$ zur Berechnung der Fakultät herangezogen, da sie als ihr reellwertiges Komplement gilt [271; 272]. Damit ergibt sich die Komplexität $X(m^*)$ einer Musterposition entsprechend Formel 4.9 im gleichen Intervall. Durch die Bildung der Summe oder des Durchschnitts der Einzelkomplexitäten kann die Definition auf ganze Muster übertragen werden.

$$X(z) = \frac{1}{L} \cdot \log_k \left(\frac{L!}{\prod_{i=1}^{k} z_i!} \right) \tag{4.8}$$

$$X(m^*) = \frac{1}{|m^*|} \cdot \log_k \left(\frac{|m^*|!}{\prod_{i=1}^{k} \Gamma\left(\widetilde{h}_i^{m^*} + 1 \right)} \right) \tag{4.9}$$

4.1.3 Bewertung einzelner Musterinstanzen

Das Ergebnis einer Mustersuche besteht aus einer großen Zahl sogenannter konkreter Musterinstanzen $W = (w_1, \ldots, w_n), \forall i : w_i \in m_i^*$. Abhängig vom Aufbau des gesuchten Musters und der Verteilung der Variabilität unter den Suchergebnissen kann es erhebliche Unterschiede in der Abbildungsgüte der Musterinstanzen geben. Für die Bewertung der Suchergebnisse im Detail ist es ist daher sinnvoll, auch ein Maß für die Abbildungsgüte einer Musterinstanz zu kennen.

Zu diesem Zweck kann die konkrete Musterinstanz W in Bezug zur PSSM des zugrunde liegenden Gesamtmusters gestellt werden. Nachdem für jede Position w_i der Musterinstanz W die zugehörige relative Häufigkeit $h_{m_i^*}(w_i)$ aus der PSSM bestimmt wurde, wird durch Multiplikation dieser Werte eine Schätzung der Auftretenswahrscheinlichkeit der Musterinstanz im Fund erreicht. Wie in Formel 4.10 gezeigt ist, wird diese Größe zur besseren Skalierung logarithmiert, wobei in der vergleichenden Bewertung die Wahl der Basis ψ ohne Einfluss bleibt. Soll die informationelle Struktur des Musterfundes in die Bewertung einfließen, so kann das Bewertungsverfahren entsprechend angepasst werden. Dazu wird für jede Musterposition der partielle Informationsgehalt $I(m_i^*)$ bestimmt und mit relativen Häufigkeiten des tatsächlich in der Musterinstanz auftretenden Elements gewichtet. Die gewichteten Informationsgehalte können schließlich entsprechend Formel 4.11 zur Gesamtbewertung aufsummiert werden. Eine Logarithmierung ist dabei nicht notwendig, da die partiellen Informationsgehalte bereits logarithmisch vorliegen.

$$S_{m^*}^h(W) = \log_\psi \left(\prod_{i=1}^{n} h_{m_i^*}(w_i) \right) \tag{4.10}$$

$$S_{m^*}^I(W) = \sum_{i=1}^{n} \left(h_{m_i^*}(w_i) \cdot I(m_i^*) \right) \tag{4.11}$$

4.1.4 Visualisierung

Zwar kann ein Muster entsprechend der obigen Definition in einer geeigneten Datenstruktur gespeichert werden, diese ist aber nur maschinell verarbeitbar. Häufig werden jedoch auch Darstellungsformen gebraucht, die von einem menschlichen Benutzer intuitiv verstanden werden können. Dies schließt sowohl die tabellarische als auch graphische Ausgabe von Sequenzmustern ein, die eine wichtige Schnittstelle zwischen algorithmischer Verarbeitung und manueller Auswertung darstellt.

Graphische Darstellung basierend auf der PSSM Basierend auf den konkreten Häufigkeitsverteilungen der PSSM kann eine graphische Darstellung des Musterfundes erreicht werden. Die beste Repräsentation wird dabei durch die Form des Sequenzlogos erreicht, das sowohl die Häufigkeitsverteilung als auch die Informationsgehalte der Musterpositionen berücksichtigt. Im Sequenzlogo ordnet sich der horizontalen Aneinanderreihung der Musterpositionen eine vertikale Staffelung der jeweils in der Position vorhandenen Elemente unter. Die Höhe einer jeden vertikalen Staffelung skaliert mit dem Informationsgehalt der korrespondierenden Musterposition und verteilt sich im Verhältnis ihrer Auftretenshäufigkeiten auf die beteiligten Elemente [273]. Eine Farbkodierung der Elemente des Alphabets trägt zur schnellen Übersicht bei. Abweichend von der eigentlichen Definition werden Lücken im weiteren durch einen schwarzen Punkt in der Mitte der Zeilenhöhe dargestellt. Ein kurzes Musterbeispiel befindet sich in Abbildung 4.2.

$$\begin{pmatrix} A \\ C \\ G \\ T \end{pmatrix} \rightarrow \begin{pmatrix} 0{,}00 & 0{,}25 & 0{,}24 & 0{,}55 & 0{,}00 & 0{,}29 & 0{,}00 \\ 1{,}00 & 0{,}25 & 0{,}28 & 0{,}00 & 0{,}00 & 0{,}24 & 0{,}00 \\ 0{,}00 & 0{,}25 & 0{,}26 & 0{,}00 & 1{,}00 & 0{,}26 & 1{,}00 \\ 0{,}00 & 0{,}25 & 0{,}22 & 0{,}44 & 0{,}00 & 0{,}21 & 0{,}00 \end{pmatrix}$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

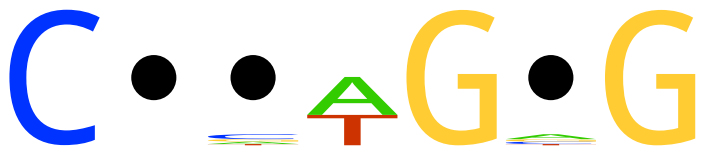


Abb. 4.2: Beispielhafte Darstellung eines kurzen, konkreten DNA-Sequenzmotivs. Die erweiterte Sequenzlogodarstellung im unteren Bereich korrespondiert direkt mit der darüber befindlichen PSSM. Die Verteilung der Häufigkeiten pro Musterposition bestimmt dabei die relativen vertikalen Größenanteile der Symbole im Sequenzlogo, welche jedoch insgesamt durch den Informationsgehalt der Musterposition skaliert wird. Der Einfluss der Häufigkeitsverteilungen auf den Informationsgehalt wird im Vergleich der Positionen deutlich. Zudem wird die neu eingeführte Darstellung von Positionen mit sehr geringem Informationsgehalt als Lücken (schwarzer Punkt) demonstriert.

Tabellarische Darstellung basierend auf der PSPM Basierend auf den Zugehörigkeitsinformationen der PSPM kann eine textuelle Darstellung des Musterfundes erreicht werden, die sich zur tabellarischen Nutzung eignet. Als Mittel der Darstellung werden reguläre Ausdrücke verwendet. Diese durch reguläre Grammatiken erzeugten, formalen Sprachen bieten ein mächtiges Werkzeug in der komplexen Zeichenkettensuche [274]. Die Komplexität des vollen Sprachumfangs beeinträchtigt jedoch rasch die intuitive Lese- und Schreibbarkeit der erzeugten regulären Ausdrücke. Zur Realisierung der Muster entsprechend der obigen Definition sind jedoch nur einfache syntaktische Konstrukte notwendig. Mit der Einschränkung auf diese Basiskonzepte bleibt auch die intuitive Nutzung der resultierenden regulären Ausdrücke erhalten. Für triviale Musterpositionen reicht die unveränderte Schreibung des Zeichenliterals bereits aus. Die Darstellung komplexer Musterpositionen erfordert eine mit eckigen Klammern eingeschlossene Auswahl aller enthaltener Zeichenliterale. So ergibt sich für das Muster in Abbildung 4.2 der reguläre Ausdruck `C[ACGT][ACGT][AT]G[ACGT]G`.

4.2 Algorithmus zur Mustersuche

Die Mustersuche mit variablen Musterpositionen gehört zur Komplexitätsklasse NP. Lösungsstrategien, die keine heuristischen Verfahren nutzen, sondern die optimale Lösung anstreben, weisen daher eine Laufzeit auf, die exponentiell zur Länge der Muster und zum Grad der Variabilität skaliert. Eine effiziente Implementierung ist folglich essentiell für die Durchführung der Suche. Prinzipiell existieren zwei unterschiedliche Herangehensweisen für die Mustersuche in Sequenzdatensätzen, die im folgenden kurz vorgestellt werden.

Die erste dieser beiden ist die direkte Ableitung von ähnlichen oder gleichen Subsequenzen aus dem Sequenzdatensatz. Dies kann durch unterschiedliche Strategien erreicht werden. Über die Lösung des ebenfalls NP-schweren Problems der längsten gemeinsamen Teilsequenz kann eine unterbrochene Folge von Zeichen gefunden werden, die in allen Sequenzen vorkommt [275].

Das Verfahren unterstützt keine Variabilität unter den Musterpositionen und ist damit intolerant gegenüber sequenziellen Mutationen, welche dazu führen, dass einzelne Sequenzen die Reinform des Musters nicht mehr enthalten. Ferner verteilt sich die gefundene längste gemeinsame Teilsequenz meist ungleichmäßig über die Sequenzen, was ihre Aussagekraft zusätzlich infrage stellt. Eine weitere algorithmische Lösungsstrategie schafft über die Durchführung multipler Sequenzalignments ebenfalls eine Möglichkeit der direkten Ableitung gemeinsamer Sequenzmuster. Für die Durchführung multipler Sequenzalignments haben sich progressive Heuristiken entwickelt, welche über die einfach bestimmbaren, paarweisen Sequenzalignments einen phylogenetischen Baum konstruieren. Anhand dieser auch *Guide Tree* genannten Baumstruktur werden die einzelnen Sequenzen schließlich schrittweise dem Alignment hinzugefügt. Lücken können dabei nicht wieder aus dem Alignment entfernt werden. Durch die verwendeten Heuristiken erreichen die Alignmentverfahren meist akzeptable Laufzeiten [276]. Aus dem fertigen Alignment können schließlich in Bereichen mäßiger bis hoher Konservierung Sequenzmuster abgeleitet werden, die dank der Toleranz des Verfahrens gegenüber Fehlpaarungen variable Positionen und Lücken enthalten können. Die Güte der eingesetzten Heuristik und ihrer Parametrisierung ist jedoch ausschlaggebend für die Qualität des Ergebnisses. Es ist jedoch bekannt, dass die Alignmentverfahren eine nicht unerhebliche Fehleranfälligkeit aufweisen und ihre Parametrisierung eine starke Fallbezogenheit aufweist. Sie bilden damit einen potentiellen algorithmischen Flaschenhals [277; 278]. Da es kein universelles Benchmark für multiple Sequenzalignments gibt [279], ist eine Validierung des Modells ohne Referenzdaten nicht möglich.

Die zweite prinzipielle Herangehensweise nutzt gängige Suchverfahren für Zeichenketten, um den gesamten Musterraum systematisch auf Funde im gegebenen Sequenzdatensatz zu untersuchen. Neben der naiven Suche in Zeichenketten bietet sich hier gerade bei großen Sequenzdatensätzen die Datenstruktur des Suffixbaumes an. Die Mehrdeutigkeit muss bei einer derartigen Suche über die Erweiterung des Musterraumes erreicht werden, bedingt jedoch eine exponentielle Vergrößerung desselben. Zur effektiven Suche ist daher zusätzlich eine Einschränkung dieses Suchraumes erforderlich. In diesem Abschnitt soll daher ein Suchverfahren entwickelt werden, welches in der Lage ist, auf Basis von Suffixbäumen eine Mustersuche in großen Sequenzdatensätzen durchzuführen. Die Suche beschränkt sich dabei im ersten Teil auf triviale Muster und wird schließlich für die Nutzung komplexer Muster um das Prinzip der Variabilität erweitert. Die dazu notwendigen, grundlegenden Arbeitsschritte sind zur besseren Übersicht in Abbildung 4.3 als Ablaufschema dargestellt.

4.2.1 Suche in Suffixbäumen

Ein Suffix einer Sequenz S ist eine endliche Teilfolge U genau dann, wenn eine weitere Teilfolge V existiert, sodass $S = V \circ U$. Eingeschlossen der leeren Teilfolge besitzt eine Sequenz S damit genau $|S|$ Suffixe. Im ursprünglichen Sinne ist ein zur Sequenz S zugehöriger Suffixbaum $T = (R_T, N_T, E_T)$ ein von Wurzelknoten R_T ausgehender, gerichteter Baum mit genau $|S|$ Blättern, dessen Kanten E_T mit jeweils einer Teilsequenz von S beschriftet sind. Jeder innere Knoten von T hat dabei mindestens zwei Kinder, deren Kantenbeschriftungen mit unterschiedlichen Elementen des Alphabets beginnen. Jedem Blatt des Baumes wird ferner genau der Suffix zugeordnet, der sich ergibt, wenn die Beschriftungen entlang des Pfades von der Wurzel bis zum betrachteten Blatt konkateniert werden. Existiert nicht für jeden Suffix ein Blatt im Suffixbaum, so spricht man von einem impliziten Suffixbaum [280; 281]. Mithilfe zusätzlicher Verknüpfungen zwischen den Knoten N_T des Baumes kann ein solcher Baum aus einer Sequenz in linearer Zeit erzeugt werden [280]. Enthält ein Suffixbaum mehrere Ursprungssequenzen, so wird er

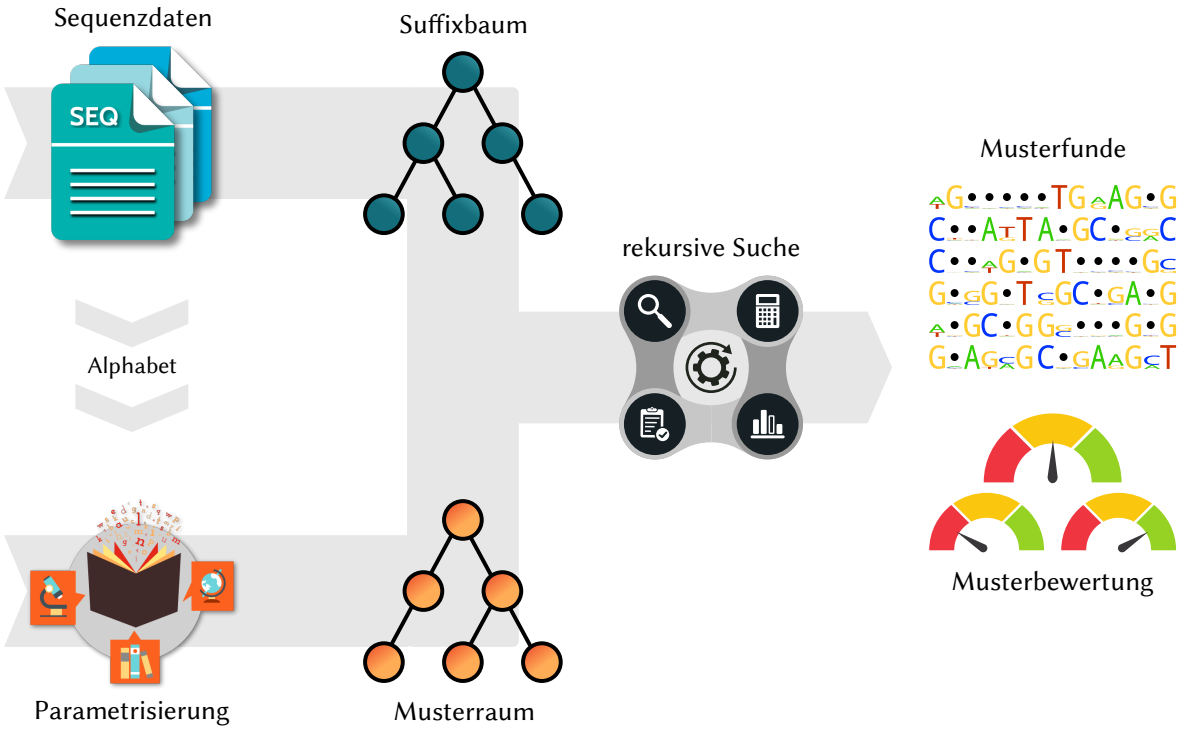


Abb. 4.3: Vorgeschlagenes Ablaufschema einer Mustersuche. Nach der Überführung der Sequenzdaten in einen Suffixbaum und der prinzipiellen Definition des Musterraumes aus der gegebenen Parametrisierung wird der rekursive Suchalgorithmus angewendet, welcher anhand der Filterkriterien bewertete Musterfunde ausgibt.

gemeinhin als generalisierter Suffixbaum bezeichnet [282]. Die Zuordnung der Suffixe in den Blättern muss nun zusätzlich auf die Ursprungssequenzen verweisen. Für die schnelle Suche nach Teilsequenzen reicht ein impliziter Suffixbaum aus. Zur genauen Zuordnung der gefundenen Teilsequenzen empfiehlt sich jedoch die zusätzliche Annotation aller Knoten mit den Lageinformationen aller durch ihn abgedeckten Subsequenzen. Die Suche einer Teilsequenz ist durch die Traversierung des Baumes nun in linearer Zeit bezogen auf die Länge der zu suchende Teilsequenz möglich [281]. Durch diese hervorragenden Laufzeiteigenschaften der Suffixbäume haben sie in den letzten Jahrzehnten in der Zeichenkettenverarbeitung eine große Bedeutung gewonnen [283–288].

Konstruktion eines Suffixbaumes Bei dem hier beschriebenen Verfahren der Mustersuche kommt ein beschnittener, generalisierter, impliziter Suffixbaum zum Einsatz. Um die Suffixe auf die Pfade des Baumes zu projizieren, ist jede Kante mit genau einem Element des Alphabets beschriftet. Da der Suffixbaum damit nicht für die Aufnahme variabler oder unbekannter Elemente konzipiert ist, müssen im Vorhinein alle Sequenzen aus dem Datensatz entfernt werden, die derartige Positionen enthalten. Zur schnelleren internen Verarbeitung wurden das Alphabet auf eine Menge ganzer Zahlen abgebildet. Auf diese Weise kann die Suche nach einem Folgeknoten in der Implementierung über Arrays realisiert werden, die einen wahlfreien und latenzarmen Zugriff ermöglichen. Zur Erfassung mehrerer Ursprungssequenzen erfordert die Datenstruktur die Speicherung der zugehörigen Referenzen. Dies wird erreicht, indem jeder Knoten des Suffixbaumes eine Menge von Referenzen auf diejenigen Sequenzen führt, welche die vom betrachteten Knoten repräsentierte Teilsequenz beinhalten. Der Wurzelknoten repräsentiert die leere Sequenz und enthält damit Referenzen auf alle Sequenzen, die in den Suffixbaum eingefügt wurden.

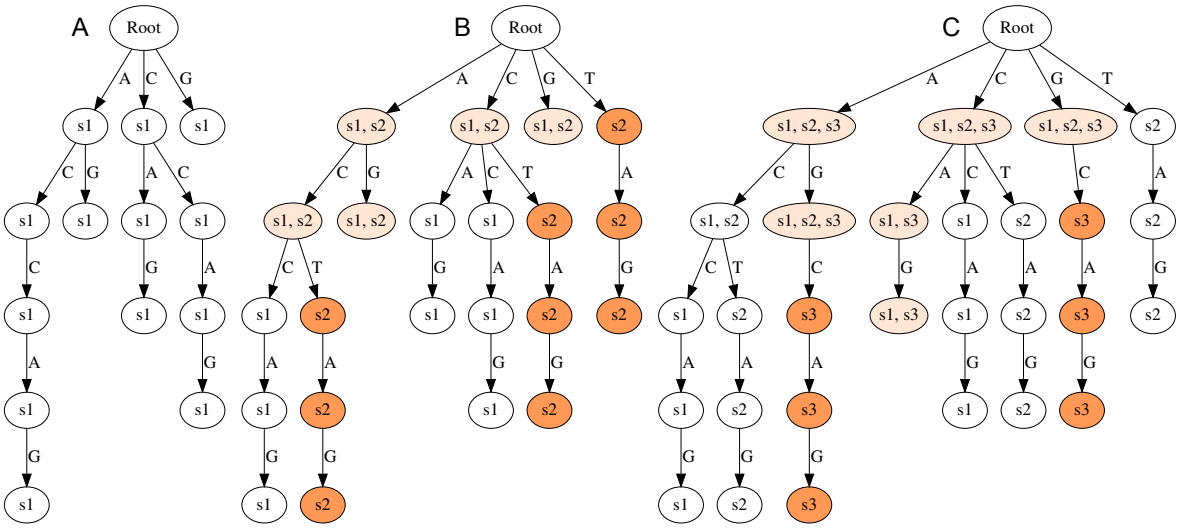


Abb. 4.4: Gezeigt wird die beispielhafte Konstruktion eines generalisierten, impliziten Suffixbaumes. In den drei Schritten A, B und C wird jeweils die Sequenz s1 (`ACCAG`), s2 (`ACTAG`) und s3 (`AGCAG`) in den Baum hinzugefügt. In der wachsenden Baumstruktur sind die in einem Schritt neu erzeugte Knoten orange und die veränderten Knoten beige eingefärbt. Die Kanten sind mit den entsprechenden Elementen des Alphabets annotiert, die konkateniert die Suffixe ergeben. Die Knoten enthalten die Sequenzliste.

Der Suffixbaum entsteht schließlich sukzessive durch das schrittweise Hinzufügen aller Suffixe der vorgegebenen Sequenzen. Dazu werden in einem vorbereitenden Schritt die Suffixe aller Sequenzen erzeugt und in einer gemeinsamen Liste hinterlegt. Für jeden Eintrag dieser Liste wird nun der Einfügeprozess nach dem folgenden Schema angestoßen. Beginnend mit dem Wurzelknoten des Baumes wird dieser elementweise anhand der einzufügenden Suffixsequenz abwärts durchschritten. Dabei wird jeweils die Kante gewählt, welche die aktuelle Sequenzposition als Beschriftung trägt. Ist diese noch nicht im Baum vorhanden, so wird sie vor dem Fortfahren neu angelegt. Jeder Knoten, der auf diesem Weg besucht wird, speichert einen Verweis auf die Sequenz, zu der der aktuell bearbeitete Suffix gehört. Da die Referenzen in einer Menge vorgehalten werden, kommt es an dieser Stelle nicht zur mehrfachen Speicherung identischer Sequenzen. Die Anzahl der Schritte dieser aktiv einfügenden Traversierung des Suffixbaumes ist durch die maximale Länge der Muster begrenzt, die später anhand der Datenstruktur gefunden werden sollen. Für einen kleinen Datensatz kurzer DNA-Sequenzen wird der Konstruktionsvorgang in Abbildung 4.4 exemplarisch durchgeführt.

Der Speicherbedarf eines solchen Suffixbaumes wächst am Beginn des Einfügevorgangs am stärksten, da in dieser Phase sowohl Knoten, Kanten als auch Referenzmengen angelegt und befüllt werden. Im späteren Verlauf sind bereits viele Teilsequenzen der Suffixe als Knoten im Baum enthalten, sodass nur noch eine Erweiterung der Referenzmengen erforderlich ist. Durch den Beschnitt des Baumes auf die Länge der später damit zu suchenden Muster wird die kombinatorische Vielfalt und damit der mögliche Speicherbedarf der Datenstruktur effektiv und ohne Verlust seiner Funktion begrenzt. Der beschriebene Suffixbaum weist damit für n Sequenzen der Länge l über dem Alphabet der Länge k und einer maximale Musterlänge von m eine maximale Speicherkomplexität von $\mathcal{O}(n \cdot k^{m+1})$ auf. Der Konstruktionsprozess kann eine Zeitkomplexität von $\mathcal{O}(m \cdot l \cdot n)$ zugeordnet werden. Da der Aufbau des Suffixbaumes verglichen mit dem späteren Suchvorgang zeitlich nur eine untergeordnete Rolle spielt, wurde auf eine Implementierung der linearen Erzeugungsstrategie vorerst verzichtet. Sie kann jedoch bei eintretender Notwendigkeit nachträglich hinzugefügt werden.

Triviale Suche in einem Suffixbaum Nachdem der Suffixbaum nach dem beschriebenen Prinzip erstellt wurde, kann eine beliebige triviale Mustersequenz eingeschränkt durch die maximal zugelassene Länge in linearer Zeit gesucht werden. Dazu wird der Baum beginnend mit seinem Wurzelknoten elementweise anhand der gesuchten Mustersequenz abwärts durchschritten. Dabei wird in jedem Schritt diejenige Kante verfolgt, welche die aktuelle Musterposition als Beschriftung trägt. Existiert diese Kante nicht, so ist das Suchmuster nicht in den Sequenzen enthalten, aus denen der Suffixbaum erzeugt wurde. Mit erreichen der letzten Musterposition gilt das Muster als gefunden. Der dabei erreichte Endknoten enthält die Referenzen zu allen Sequenzen, die das triviale Muster enthalten. Zum Suchen komplexer Muster in einem solchen Suffixbaum muss das Verfahren jedoch algorithmisch erweitert werden. Mit der Einführung des *Pattern Discovery Algorithm* wurden drei mögliche Motivklassen festgelegt, von denen eine als Motive mit Zeichengruppen der Definition aus Abschnitt 4.1 weitgehend entspricht. Der vorgestellte Algorithmus nutzt einen Suffixbaum, dessen Kanten mit Zeichengruppen beschriftet sind [289]. Durch diese Erweiterung des Alphabets ist es möglich, komplexe Muster im Suffixbaum abzubilden und anschließend über eine triviale Suche zu finden. Die Größe des Suffixbaumes und der erforderliche Aufwand bei dessen Erzeugung wird dabei jedoch derart stark erhöht, dass die Datenstruktur speichertechnisch selbst unter Verwendung optimierter Konstruktionsalgorithmen nicht mehr tragbar ist.

Komplexe Suche durch progressives Node Merging Da die inhaltliche Variabilität der Musterpositionen nicht über die Datenstruktur realisiert werden kann, wird eine algorithmische Lösung vorgeschlagen. Die dabei verfolgte Strategie nutzt das Prinzip der progressiven horizontalen Zusammenführung der relevanten Knoten des Baumes, welches im Folgenden auch kurz progressives *Node Merging* genannt wird. Gegeben sind ein komplexes Sequenzmuster M^* der Länge n sowie ein generalisierter, impliziter Suffixbaum T über dem gleichen Alphabet. Für die Mustersuche wird ferner eine interne Knotenmenge $\mathcal{K} \subseteq N_T$ definiert. Zur Zwischenspeicherung des bisher bearbeiteten Teilmusters $\mathcal{M}_i$, einem Präfix variabler Länge i des Suchmusters M^*, wird eine weitere Größe bereitgehalten. Zu Beginn einer komplexen Suche existiert eine initiale Knotenmenge $\mathcal{K}_0 = \{R_T\}$, die nur den Wurzelknoten des Suffixbaumes enthält. Das korrespondierende, bisher erreichte Teilmuster $\mathcal{M}_0$ ist leer. Für jede Musterposition m_i^* wird nun der folgende Ablauf konsekutiv wiederholt. Das bisher bearbeitete Muster wird entsprechend Formel 4.12 um die aktuelle Musterposition m_i^* erweitert. Für jeden Knoten k der vorigen Knotenmenge $\mathcal{K}_{i-1}$ werden alle Folgeknoten $F_T(k, m_i^*)$ nach Formel 4.13 bestimmt, deren zugehörige Kanten durch die Elemente der aktuellen Musterposition m_i^* beschriftet sind. Schließlich wird die aktuelle Knotenmenge $\mathcal{K}_i$ entsprechend Formel 4.14 als Vereinigungsmenge dieser Folgeknoten gebildet. Aus der finalen Knotenmenge $\mathcal{K}_n$ kann nun die Menge aller Sequenzen bestimmt werden, die das gegebene Muster M^* beinhalten. Mehrfachvorkommen einzelner Sequenzen werden durch die Mengendefinition bereits im Vereinigungsprozess unterbunden. Das Prinzip des progressiven *Node Merging* wird in Abbildung 4.5 an zwei einfachen Sequenzen verdeutlicht.

$$\mathcal{M}_i = \mathcal{M}_{i-1} \circ m_i^* \tag{4.12}$$

$$F_T(k, m^*) = \{x \in N_T : \exists m \in m^* \wedge (k,x,m) \in E_T\} \tag{4.13}$$

$$\mathcal{K}_i = \bigcup_{k \in \mathcal{K}_{i-1}} F_T\,(k,m_i^*) \tag{4.14}$$

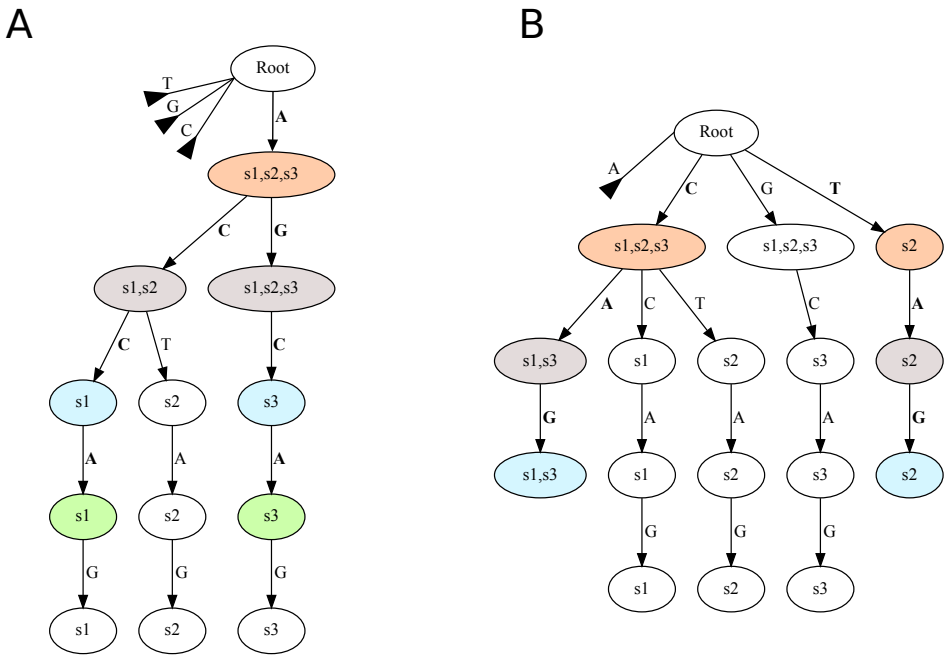


Abb. 4.5: Die Suche eines komplexen Musters nach dem Prinzip des *Node Merging* wird für zwei einfache Beispielmuster demonstriert. Es wird in beiden Fällen der Suffixbaum verwendet, der in Abbildung 4.4 erzeugt wurde. Nichtrelevante Äste wurden durch ein schwarzes Dreieck ausgeblendet. Die farbliche Kennzeichnung entspricht den aufeinanderfolgenden Suchschritten von orange nach grün. Die Beschriftungen von Kanten, die im Suchverlauf gewählt wurden, sind fett gedruckt. Durch die Suche des Musters `A[CG]CA` auf der linken Seite ergeben sich nach einer initialen Aufspaltung im zweiten Suchschritt schließlich zwei Knoten (grün), aus denen die Sequenzmenge $\{s1, s3\}$ abgeleitet werden kann. Analog ergibt die Suche nach dem Motiv `[CT]A[GT]` auf der rechten Seite die Sequenzmenge $\{s1, s2, s3\}$ (cyan).

4.2.2 Durchsuchen des Musterraumes

Neben der Suche eines bestimmten Musters liegt ein weiteres Hauptaugenmerk für die Entwicklung des Mustersuchverfahrens auf der systematischen Überprüfung des gesamten möglichen Musterraumes auf Vertreter, die im gegebenen Sequenzdatensatz überrepräsentiert vorliegen. Nachdem der Sequenzdatensatz wie beschrieben in einen Suffixbaum überführt wurde, erfolgt die Zusammenstellung der Menge aller erlaubten komplexen Musterpositionen $\mathfrak{M}$. Über die maximal erlaubte Anzahl von Elementen c einer solchen komplexen Musterposition kann der Grad der Variabilität und damit die Größe des Suchraumes eingeschränkt werden. Um trotz dieser Einschränkung Lücken in den Mustern zuzulassen, bietet sich die Erweiterung $\mathfrak{M}_+$ an. Durch diese wird zusätzlich die vollständig unbestimmte Musterposition $m^*_\circ$ zugelassen, welche bei eingeschränkter Variabilität ansonsten grundsätzlich entfällt. Beide Mengendefinitionen befinden sich in Formel 4.15. Der Suchraum $\mathfrak{S}$ umfasst schließlich alle komplexen Muster, die unter der spezifizierten Parametrisierung zugelassen sind. Durch Festlegung der minimalen l_{min} und maximalen l_{max} erlaubten Musterlänge kann der Suchraum wie in Formel 4.16 zusammengefügt werden. Die Größe des Musterraumes wächst exponentiell mit der erlaubten Musterlänge l_{max}.

$$\mathfrak{M} = \{m^* : m^* \subseteq A \wedge |m^*| \leq c\} \qquad \mathfrak{M}_+ = \mathfrak{M} \cup \{A\} \tag{4.15}$$

$$\mathfrak{S} = \bigcup_{i=l_{\text{min}}}^{l_{\text{max}}} \left(\mathfrak{M}_{(+)}\right)^i \tag{4.16}$$

Rekursive Suchstrategie Bereits bei den üblich verwendeten Musterlängen lässt der große Umfang des Suchraumes keine Speicherhaltung aller möglichen Muster mehr zu. Um ihn vollständig und systematisch auf Funde zu prüfen, wird daher eine rekursive, erschöpfende Strategie angewendet. Die Kerneinheit dieser Strategie wird im folgenden als rekursive, algorithmische Einheit bezeichnet. Sie erfordert ein Ausgangsmuster $\mathcal{M}_i$ als einzigen Parameter. Für dieses Muster wird über das bekannte Verfahren des *Node Merging* basierend auf dem gegebenen Suffixbaum die Liste aller Sequenzen bestimmt, die das Muster enthalten. Neben seiner Vorkommenshäufigkeit wird aus den konkreten Musterfunden die PSSM-Darstellung des Musters abgeleitet. Basierend auf diesen Daten entscheidet ein benutzerdefiniertes Kriterium über die Annahme des Musters als gültigen Fund. Sofern die maximale Musterlänge $l_{\max}$ noch nicht erreicht wurde, wird anschließend für jede erlaubte Musterposition $m^* \in \mathfrak{M}_{(+)}$ aufeinanderfolgend eine weitere rekursive Einheit aufgerufen, welche die Konkatenation aus Ausgangsmuster und neuer Musterposition $\mathcal{M}_{i+1} = \mathcal{M}_i \circ m^*$ erhält. Die Rekursion bricht daher die realisierte Tiefensuchstrategie nach Erreichen der maximalen Länge selbstständig ab. Der initiale Aufruf der rekursiven, algorithmischen Einheit erhält das leere Muster als Parameter. Ohne weitere Bedingungen garantiert das Verfahren die vollständige Prüfung aller infrage kommenden Muster des vorgestellten Suchraumes $\mathfrak{S}$.

Prüfung der Muster Das benutzerdefinierte Kriterium, welches über die Annahme eines Musters entscheidet, prüft neben der Länge des Musters die festgestellte Vorkommenshäufigkeit relativ zur Größe des Sequenzdatensatzes. Die Überprüfung dieser beiden Kriterien erfolgt in sehr kurzer Zeit, da die notwendigen Größen ohne weitere Berechnungen zur Verfügung stehen. Für alle weiteren Kriterien ist die konkrete PSSM des Musters notwendig, welche durch Auswertung aller beteiligten Sequenzen abgeleitet werden muss. Die Instanziierung der PSSM verursacht durch die Extraktion der Detailinformationen aus dem Sequenzdatensatz einen nicht unerheblichen, zusätzlichen Berechnungsaufwand.

Mit ihrer Hilfe kann jedoch die Integrität eines Musters bestimmt werden. Ein Muster gilt dann als integer, wenn die vorgegebene Variabilität aller Musterpositionen tatsächlich mit der im Fund beobachteten Variabilität übereinstimmt oder kein weiteres Muster im Suchraum existiert, welches die Variabilität des Musters genauer abdeckt. Für jede komplexe Musterposition m^* wird dazu aus der PSSM die Menge aller tatsächlich vorkommenden Elemente μ^* abgeleitet. Falls die tatsächliche Variabilität $\mu^* \subset m^*$ kleiner ist als im Muster beschrieben, so ist zu überprüfen, ob eine andere erlaubte Musterposition $m^*_< \in \mathfrak{M}_{(+)}$ die Fundstelle mit geringerer Variabilität $\mu^* \subseteq m^*_< \subset m^*$ genauer beschreiben könnte. Ist dies der Fall, so existiert ein weiteres Muster im definierten Suchraum $\mathfrak{S}$, welches den Musterfund genauer beschreibt. Das aktuell betrachtete Muster kann folglich verworfen werden.

Auf Basis der aufgestellten Bewertungsmaße für Muster können weitere Kriterien für die benutzerdefinierte Ergebnisfilterung definiert werden. Für die direkte Auswertung innerhalb des Suchprozesses eignet sich besonders der Informationsgehalt. Er bietet eine bewährte numerische Darstellung der tatsächlichen Variabilität und ist mit geringem Aufwand berechenbar. Trotz Vorstufen der Filterung ist davon auszugehen, dass ein großer Anteil des Suchraumes durch dieses Kriterium geprüft werden muss. Bei der Vielzahl der erwarteten Prüfungen wird der Einsatz der anderen Maße aufgrund ihres im Vergleich höheren Berechnungsaufwandes nur nachrangig möglich sein. Das Filterkriterium lässt eine Einschränkung des Informationsgehalts pro Musterposition, aber auch über den Durchschnitt des gesamten Musters zu. Es ist damit ein essentielles Werkzeug zur Regulation der Aussagekraft der zu erwartenden Musterfunde. Ohne

die Anwendung eines solchen Filters ist eine sehr große Anzahl von Musterfunden zu erwarten, die aufgrund ihrer hauptsächlichen Zusammensetzung aus hochvariablen Musterpositionen bedeutungslos sind und die Interpretation der Suchergebnisse nahezu unmöglich machen. Einzige Ausnahme bilden an dieser Stelle einzeln auftretende Lücken. Unter Zuhilfenahme eines Schwellwertes können einzelne Musterpositionen mit besonders niedrigem Informationsgehalt als Lücken markiert werden. Diese unterliegen nicht der normalen Bewertung nach dem Informationskriterium. Da sie nur in der Mitte eines Musters sinnhaft sind, können Muster mit randständigen Lücken prinzipiell verworfen werden.

Integration von rekursiver Suche und Node Merging Während die rekursive Suchstrategie den baumförmigen Suchraum traversiert, arbeitet des Prinzip des *Node Merging* mit dem Suffixbaum. Auf den ersten Blick unvereinbar haben die beiden Verfahren jedoch eine wichtige Gemeinsamkeit. Sie arbeiten nach dem Prinzip der Tiefensuche. Nach dem Abstieg der rekursiven Suchstrategie in die nächsttiefere Rekursionsebene teilen alle Anfragen für die Mustersuche nach dem Prinzip des *Node Merging* den gleichen Musterpräfix. Entsprechend teilen diese Anfragen für den Bereich dieses Präfixes den selben Abstieg während des *Node Merging*. Auf dieser Basis ist eine Wiederverwendung der suchinternen Knotenmenge möglich. Die erforderlichen Rückschritte im Suchraum bei Erreichen eines Abbruchkriteriums sind im *Node Merging* jedoch nicht ohne weiteres abbildbar. Ihre Realisierung erfordert die Zwischenspeicherung der Knotenmengen in jeder Rekursionsebene des Abstiegs. Da dies jedoch nur für den jeweils Aktuellen Suchstrang notwendig ist, muss durch die Erweiterung kein Speicherengpass erwartet werden. Die algorithmischen Kerne der beiden bisher getrennten Verfahren wurden für die Integration aufgebrochen und miteinander verbunden. Die direkte Verwaltung der Knotenmenge in der rekursiven, algorithmischen Einheit verringert auf diese Weise den notwendigen Aufwand eines einzelnen Mustersuchvorgangs auf nahezu konstante zeitliche Komplexität.

4.2.3 Optimierung der Suchstrategie

Zur Optimierung der Suchstrategie werden die Prinzipien des *Branch and Bound*-Verfahrens zur Einschränkung des Suchraumes auf die bisherige Strategie angewendet. Das Meta-Verfahren *Branch and Bound* wurde erstmals im Bereich der Unternehmensforschung beschrieben [290]. Es wird auf kombinatorische Optimierungsprobleme angewendet, welche ansonsten der vollständigen Durchmusterung eines komplexen, großen Suchraumes bedürfen. Das grundlegende *Branch and Bound*-Verfahren erfordert die Überführung des Suchraumes in eine baumartige Struktur, was eine entsprechende natürliche Topologie des Suchraumes voraussetzt. Die Verringerung des Suchraumes stützt sich auf spezielle Ausschlusskriterien, auch Schranken genannt, die nicht nur für ein einziges Element des Suchbaumes gültig sind, sondern für ganze Teilbäume. Üblicherweise kann davon ausgegangen werden, dass mit zunehmender Tiefe im Suchbaum strengere Schranken definiert werden können. Die Durchmusterung erfolgt nun entlang der Baumstruktur des Suchraumes derart, dass nicht nur einzelne Lösungen sondern auch ganze Teilbäume über die Schranken ausgeschlossen werden. Der Umfang dieser Ausschlüsse ist stark von der Strenge der eingesetzten Schranken abhängig und kann bei Optimierungsproblemen progressiv angepasst werden. Auch die zum Suchzeitpunkt beste Lösung kann eine solche Schranke darstellen. Auf diese Weise reduziert sich die effektive Größe des Suchraumes, ohne infrage kommende Lösungen zu verlieren, wie es bei heuristischen Strategien der Fall ist. Im ungünstigsten Fall muss der gesamte Suchraum durchsucht werden.

Anwendung des Branch and Bound-Prinzips Bei dem vorgestellten Problem der Mustersuche handelt es sich nicht um ein Optimierungsproblem im eigentlichen Sinne, sodass es nicht möglich ist, aus der bisherig besten Lösung eine weitere, progressive Schranke für die folgende Suche abzuleiten. Es lässt sich jedoch über die Formulierung eines einfachen Optimierungskriteriums auf ein Optimierungsproblem mit vielen gleichwertigen Lösungen zurückführen. Dieses Kriterium beschreibt lediglich die Zugehörigkeit eines Musters zur Lösungsmenge im Wertebereich $\{0,1\}$. Das grundlegende Prinzip des *Branch and Bound* kann daher auch auf dieses Suchproblem angewendet werden, um den Musterraum effektiv zu verkleinern. Die bisher eingesetzte rekursive Tiefensuchstrategie nutzt bereits die hierarchische Topologie des Musterraumes aus, um ihn geordnet zu traversieren. Die eingesetzten Filterkriterien führen jedoch nicht zu einem Abbruch der Suche für den folgenden Teilbaum, sondern lediglich zu einer Entscheidung über den Ausschluss des untersuchten Musters. Aus den bestehenden benutzerdefinierten Kriterien werden folgend Schranken abgeleitet, welche in einer verbesserten Variante des Suchverfahrens in der Lage sind, Teilbäume des Suchraumes von der weiteren Verfolgung auszuschließen.

Nutzbare Schranken Grundlegend für die weitere Betrachtung ist die baumartige Topologie des Musterraumes. Beginnend mit dem leeren Muster in der Wurzel werden diese beim Abstieg in der Baumstruktur durch Konkatenation weiterer Musterpositionen sukzessive endständig erweitert. Durch diese Erweiterung werden die Muster zunehmend spezialisiert und die zugehörigen Fundmengen entsprechend verjüngt. Die Musterlänge erfordert als natives Abbruchkriterium der Rekursion keine separate Betrachtung, da sie selbst das Ende des Suchraumes definiert. Die ermittelten Auftretenshäufigkeiten der Muster können aufgrund der eintretenden Spezialisierung beim Abstieg im Musterraum nicht steigen. Bei Unterschreiten des eingestellten Grenzwertes kann daher der gesamte Teilbaum aus der Suche ausgeschlossen werden, ohne Gefahr zu laufen, gültige Ergebnisse zu verlieren. Ähnlich verhält es sich mit dem Kriterium der Musterintegrität. Wie bereits festgehalten, führt die zunehmende Spezialisierung der Muster beim Abstieg des Musterraumes zur Ausdünnung der Fundmenge. Durch das Entfernen einzelner Musterfunde kann die Variabilität der Musterpositionen jedoch nicht wachsen, sondern allenfalls stagnieren oder fallen. Die fehlende Integrität eines Musters kann folglich durch Spezialisierung nicht wiederhergestellt und daher als hinreichender Grund für die Auslassung des ihm angeschlossenen Teilbaums angesehen werden.

Die Bewertung der Information weist dagegen ein komplizierteres Verhalten bei der Spezialisierung eines Musters auf. Abhängig von der Verlagerung der konkreten Elementvorkommen kann der Informationsgehalt sowohl sinken als auch steigen. Die Unterschreitung der Grenzwerte für die Information einer Position oder eines ganzen Musters kann daher nicht in jedem Fall für den darunterliegenden Teilbaum angenommen werden. Als Schranke für die *Branch and Bound*-Strategie eignen sich die Informationskriterien nicht. Denkbar wäre hier nur die heuristische Variante einer Schranke, die jedoch möglicherweise Ergebnisse ungewollt von der Suche ausschließt.

4.2.4 Ordnung der Ergebnisse

Auch unter Zuhilfenahme der beschriebenen Filterkriterien resultiert die Einführung der Variabilität in einer großen Menge von Mustersuchergebnissen. Jedes dieser Ergebnisse ist mit einer Reihe von Bewertungsmaßen versehen, deren Zusammenspiel und korrekte Gewichtung nicht trivial sind. Die Fülle dieser Informationen ist nicht nur für die menschliche Interpretation

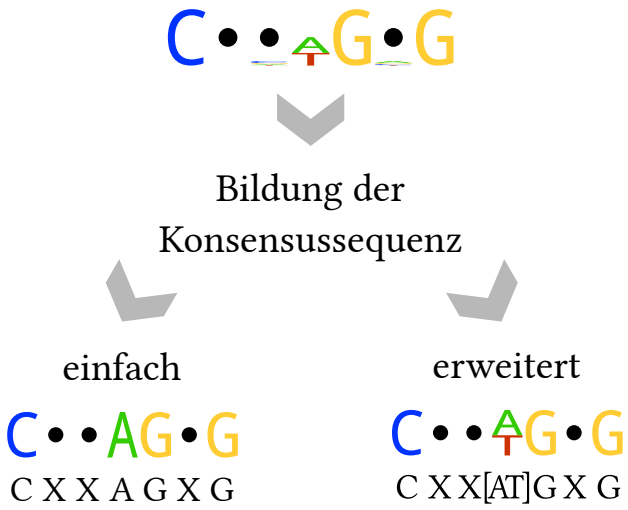


Abb. 4.6: Verringerung der Variabilität bei der Bildung der Konsensussequenz, beispielhaft gezeigt am Muster aus Abbildung 4.2. Die Bildung der einfachen Konsensussequenz (links) entfernt zugunsten einer einfachen Darstellung alle Variabilität, sodass die Information an der vierten Musterposition verloren geht. Ähnliche Muster mit abweichender Verteilung an dieser Stelle (mehr `T` als `A`) werden kontraintuitiv zu einer anderen Konsensussequenz (`CXXTGXG`) zugeordnet. In der erweiterten Konsensussequenz (rechts) bleibt die relevante Variabilität erhalten, sodass es auch bei geringer Abweichung zur korrekten Zuordnung kommt.

schwierig handzuhaben, sondern stellt auch die algorithmische Auswertung vor ein schweres Entscheidungsproblem. Unter den Musterfunden kommt es durch die Variabilität zu großen gegenseitigen Ähnlichkeiten und Überschneidungen um die tatsächlichen, intrinsischen Muster der Ausgangssequenzen. Es gilt daher, die Musterfunde in entsprechende Gruppen einzuteilen, aus denen die intrinsischen Muster dann durch Reduktion des Rauschens herausgestellt werden können.

Die einfache Konsensussequenz Ein bekannter Ansatz zur Bestimmung einer gültigen Gruppierung ist die Ableitung der Konsensussequenz. Sie stellt das triviale Muster dar, das durch Auswahl des häufigst anzutreffenden Elements $\arg\max_{a \in m^*} h_{m^*}(a)$ einer jeden Musterposition m^* erhalten wird. Beim Auftreten von Lücken ist die obige Definition zwar syntaktisch korrekt, weist aber ein semantisches Problem auf. Das häufigst anzutreffenden Element hat bei einer Lücke keine Bedeutung und würde damit die Zuordnung der Muster zu Gruppen stören. Zur Modellierung von Lücken wurde die Definition daher um ein weiteres Zeichen $X \notin A$ erweitert, welches nicht im zugrundeliegenden Alphabet vorkommt. Kommen Lücken unterschiedlicher Länge vor ohne dass diese Relevanz für die Analyse aufweist, so besteht darüber hinaus die Möglichkeit, ein weiteres Zeichen $Y \notin A$ für Lücken beliebiger Länge einzusetzen. Die einfache Konsensussequenz eignet sich zum Herausstellen trivialer intrinsischer Muster gut und deckt damit einen nicht unerheblichen Anteil bereits ab. Enthält ein intrinsisches Muster jedoch vereinzelte, bedeutungsvolle variable Positionen, so erfolgt die Gruppierung mit der einfachen Konsensussequenz wie in Abbildung 4.6 nicht optimal.

Die erweiterte Konsensussequenz Die Erweiterung der Konsensussequenz um einen gewissen Grad von Variabilität schafft hier Abhilfe. Abweichend von der ursprünglichen Definition wird zur Konstruktion einer erweiterten Konsensussequenz ein Schwellwert ϵ festgelegt, der die Inklusion der Variabilität regelt. Für jede Musterposition m^* kann anschließend die verringerte Menge der Zeichen über die Funktion $\mathrm{EK}(m^*)$ in Formel 4.17 bestimmt werden. Über die

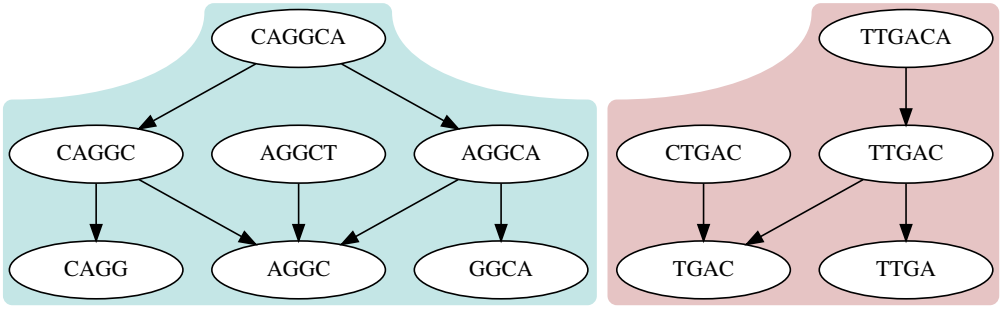


Abb. 4.7: Der Konsensusgraph zeigt die Ordnung der Konsensussequenzen durch die Enthalten-In-Beziehung beispielhaft für Suchergebnisse einer Suche im DNA-Alphabet. Der Zusammenhang der einzelnen Konsensussequenzen wird über die Kanten des gerichteten Graphen intuitiv dargestellt. Die beiden Zusammenhangskomponenten des Graphen zeigen die unabhängigen Musterfunde an (farbliche Hinterlegung).

Darstellung als regulären Ausdruck wird schließlich die erweiterte Konsensussequenz gebildet. Um die Vergleichbarkeit zu wahren, wird die Auswahl der Zeichenliterale in einheitlicher Weise sortiert. Eine Zuordnung der konkreten Musterfunde zu ihren erweiterten Konsensussequenzen liefert eine Gruppierung der Ergebnisse, die einer menschlichen Auswertung zugänglicher ist und indes relevante Variabilität berücksichtigt. Siehe dazu Abbildung 4.6.

$$\mathrm{EK}(m^*) = \{a : a \in m^* \wedge h_{m^*}(a) \geq \epsilon\} \tag{4.17}$$

Darstellung als Graph Durch das Zulassen eines Längenbereiches für Muster besteht die Möglichkeit, dass auch kürzere Teilmuster in der Ergebnismenge vertreten sind. In der Auswertung sollte dieser Tatsache Rechnung getragen werden, um den Fokus auf die wichtigen Ergebnisse zu richten. Hilfreich ist dabei die Betrachtung der Enthalten-In-Relation. Diese zusätzliche Ebene der Verbundenheit kann in der Darstellung der bereits ermittelten Konsensussequenzen durch einen gerichteter Graphen realisiert werden. Während die gefundenen Konsensussequenzen als Knoten des Graphen fungieren, wird die Enthalten-In-Relation über die Kanten modelliert. Nicht zusammengehörende Musterfunde manifestieren sich in dieser Darstellungsform durch die Ausbildung unterschiedlicher Zusammenhangskomponenten, was eine manuelle Auswertung jedoch nicht ersetzen kann. Unter Ausnutzung der Transitivität der Enthalten-In-Relation können zur Steigerung der Übersichtlichkeit alle Kanten aus dem Graphen entfernt werden, die durch indirekte Wege über andere Knoten ersetzbar sind. Ein Beispiel dafür wird in Abbildung 4.7 gezeigt.

4.2.5 Zusammenfassung

Für die Mustersuche in großen Sequenzdatensätzen wurde aufgrund der guten Eignung dieser Datenstruktur eine Lösung auf Basis von Suffixbäumen umgesetzt. Die Unterstützung für die Variabilität biologischer Sequenzen wurde über das Prinzip des *Node Merging* in den Suchalgorithmus integriert. Sie erfordert zum Erreichen eines akzeptablen zeitlichen Verhaltens moderate bis strenge Begrenzungskriterien. Da diese jedoch zum Erreichen einer hohen Güte der Suchergebnisse ohnehin notwendig sind, handelt es sich dabei nicht um eine praktische Einschränkung. Die algorithmische Umsetzung wird jedoch durch die Größe des zugrundeliegenden Alphabets und durch den Grad der erlaubten Variabilität limitiert. Für große Alphabete muss daher eine einschränkende Definition der erlaubten Variabilität vorgenommen werden.

5 Auswertung der Tertiärstruktur von Protein-Nukleinsäure-Komplexen

Die Auswertung großer Mengen an Sequenz- und Sekundärstrukturdaten ist ein wichtiges Werkzeug in der post-experimentellen Analyse von Aptameren, unterliegt jedoch den folgenden Einschränkungen. Bereits aus experimenteller Sicht ist die Menge verfügbarer Sequenzdaten meist deswegen limitiert, weil es bei der Suche nach einem passenden Aptamer in der Regel nicht zweckmäßig ist, die Selektionsbibliothek mit möglichst hoher Abdeckung zu sequenzieren, sondern lediglich die besten Vertreter daraus zu gewinnen. Eine Bestimmung der Affinität erfolgt des hohen Aufwands wegen meist nur für diese wenigen erfolgversprechenden Kandidaten. Die Anforderungen der in Kapitel 3 vorgeschlagenen Methode sind daher oft nicht oder nur unzureichend erfüllt. Auch im Falle einer ausreichenden Informationslage ist bei einer Analyse der Sequenz- und Sekundärstrukturdaten generell Vorsicht geboten. Zwar kann auf diese Weise die Charakteristik der erfassten Sequenzdaten näher beleuchtet werden, dies geschieht jedoch stets isoliert vom eigentlichen Zielmolekül. Während der experimentellen Phase ist das Zielmolekül entscheidend für die Entwicklung der Bibliothek und damit implizit für die beobachtete Charakteristik verantwortlich. Die Gewinnung zuverlässiger Informationen über den tatsächlich gebildeten Molekülkomplex, sowie den Aufbau und die Eigenschaften der molekularen Schnittstellen ist auf dieser Basis jedoch nicht möglich. Gerade diese Informationen sind jedoch essentiell für ein tiefes Verständnis der Aptamerfunktionsweise. Fundierte Aussagen über die tatsächlichen molekularen Wechselwirkungen können jedoch nur in Bezug zum Zielmolekül getroffen werden. Ohne das Vorhandensein der konkreten Bindungsgeometrie sollten die Ergebnisse der 1D- und 2D-Analyse prinzipiell nur als wichtige Hinweise gewertet werden, die im Nachgang am konkreten Komplex aus Aptamer und Zielprotein auf ihre Richtigkeit hin überprüft werden müssen. Die Aufklärung oder Simulation der Tertiärstrukturen der relevanten Aptamere im Komplex mit dem passenden Zielmolekül bilden die Grundlage dafür.

Das Zielmolekül weist als Protein im Vergleich zum Nukleotidaptamer nicht nur durch eine meist längere Molekülkette, sondern auch durch die größere Vielfalt an Grundbausteinen und möglichen Interaktionen tendenziell eine höhere strukturelle Komplexität auf. Auch wenn Primär- und Sekundärstrukturähnlichkeiten in großen Datenmengen statistisch ausgewertet werden können [291–293], finden sich die für die Bindung relevanten Charakteristika der Proteine häufig in lokalen Teilbereichen der Tertiärstruktur wieder, die nicht zwangsläufig in Sequenz oder Sekundärstruktur konserviert sind [294–296]. Derartige funktionelle Zentren können als physikochemische Interaktionsvermittler tolerant gegenüber Mutationen in der makromolekularen Ebene sein [297]. Auf Seiten der Aptamere zeigt sich zwar nur eine begrenzte Menge an Bindungsmodi der Nukleobasen, ihre Flexibilität ist jedoch nicht zuletzt durch das bewegliche Rückgrat und den Anteil unverpaarter Nukleobasen erheblich [298–300]. Bei der Selektion und Analyse von Aptameren für unterschiedliche Zielmoleküle wurden bereits mehrfach sehr spezifische Bindemotive beobachtet [301–303]. Die Ausbildung solcher Motive bedingt im Gegensatz

zu denen bei Proteinen meist das Vorhandensein lokaler, sequenziell benachbarter Strukturelemente. Diese werden in Folge der kürzeren Molekülketten und einfacheren Tertiärstrukturausformungen bei Nukleinsäuren begünstigt.

Der Grad der strukturellen Aufklärung für Proteine, Nukleinsäuren und deren Komplexe ist in den letzten Jahren sowohl in Qualität als auch in Quantität deutlich gestiegen. Daran ist nicht zuletzt die kontinuierliche Verbesserung der Aufklärungs- und Simulationsverfahren sowie die Entwicklung vielversprechender neuer Technologien wie *Cryo-Electron Microscopy* [304] beteiligt. Der Fortschritt in der Strukturaufklärung verteilt sich jedoch nicht gleichmäßig auf die genannten Molekülklassen. So liegt die Quote der Aufklärung von Protein-Nukleinsäure-Komplexstrukturen deutlich hinter der von Protein-Protein-Komplexen zurück. Bei der Betrachtung der Inhaltsstatistik der *Protein Data Bank* (PDB) [305] wird dies durch die sehr geringe Anzahl von Protein-Nukleinsäure-Strukturen von knapp unter $5\,\%$ sichtbar [306; 307]. Nur ein Bruchteil dieser Komplexstrukturen entfällt tatsächlich auf Aptamer-Komplexe. Nukleotidaptamere ohne weitere chemische Modifikationen sind auf atomarer Ebene nicht von anderen im Komplex mit Proteinen befindlichen Nukleinsäuren unterscheidbar, da physikochemisch gesehen die gleichen Grundsätze von Geometrie und Interaktion gelten. Bezogen auf die Ausbildung der Sekundärstruktur werden bei Aptameren zwar funktionsbedingt bestimmte Charakteristika bevorzugt, dies ist jedoch bei der physikochemischen Detailbetrachtung von geringer Relevanz. Bei der Analyse der Komplexstrukturen kann also zur Anreicherung der Datengrundlage auf natürliche Nukleinsäurestrukturen zurückgegriffen werden, die keine Aptamere sind.

Die experimentellen Herausforderungen und Hürden [308; 309] führen dazu, dass die Verfügbarkeit aufgeklärter 3D-Komplexstrukturen in den meisten Anwendungsfällen nicht gegeben ist. Besonders wenn die Analyse der Bindungseigenschaften einer Reihe von Aptamerkandidaten viele Strukturmodelle erfordert, wird erkennbar, dass die Strukturaufklärung kein Hochdurchsatzverfahren ist [310]. In der Praxis bietet sich daher oft nur das Ausweichen auf simulierte Komplexstrukturen an. Dazu werden die beiden Komplexpartner separat modelliert und anschließend durch eine Dockingsimulation zusammengeführt. Aus Gründen der zeitlichen Berechenbarkeit werden dazu näherungsweise Kraftfelder eingesetzt, die auf Basis einer zeitlichen und räumlichen Diskretisierung arbeiten. Die Simulation liefert dabei kein optimales Ergebnis, sondern viel mehr eine Reihe möglicher suboptimaler Kandidatenstrukturen, aus denen schließlich präferierte Kandidaten für das weitere Verfahren ausgewählt werden. Die Bewertung der Interaktionen zwischen Proteinen und Nukleinsäuren ist daher ein essentieller Bestandteil des Analyseprozesses, dem über die interne Bewertungsfunktion des eingesetzten Softwarepakets hinaus große Beachtung geschenkt werden muss.

5.1 Übersicht der Bewertungsmodelle für Protein-Nukleinsäure-Komplexe

In der Literatur finden sich sehr verschiedenartige Ansätze, die grundlegend in jene zwei Kategorien eingeteilt werden können, die in der obersten Gliederungsebene der Übersicht in Abbildung 5.1 dargestellt werden. Wissensbasierte Ansätze stellen die erste dieser Gruppen dar. Sie nutzen eine statistische Analyse bekannter Protein-Nukleinsäure-Komplexstrukturen, um entsprechende Paarpotentiale zur Bewertung abzuleiten [315]. Nicht alle bekannten wissensbasierten Ansätze finden Anwendung auf Protein-Nukleinsäure-Komplexe, so zeigten sich beispielsweise die sogenannten μ-Potentiale aus dem Bereich der Proteinfaltung als ungeeignet [316–318]. Molekularmechanische Ansätze approximieren die tatsächlichen physikalischen

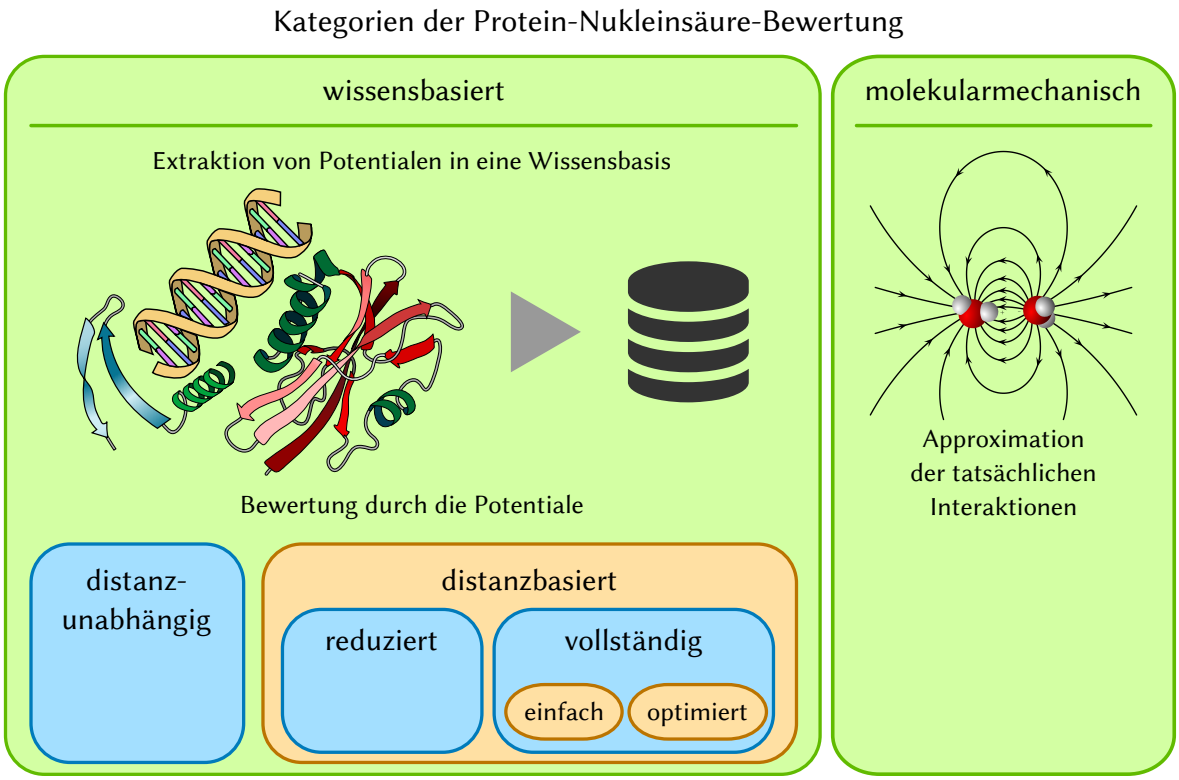


Abb. 5.1: Schematische Übersicht über die grundlegenden Beschreibungskonzepte für Protein-Nukleinsäure-Komplexe. Die zwei grundlegenden Prinzipien, die im Haupttext genannt werden, sind jeweils grün hervorgehoben. Im oberen Teil der Boxen befindet sich jeweils eine stark vereinfachte Skizze des Bewertungsprinzips. Im unteren Teil folgt gegebenenfalls eine weitere Untergliederung der Ansätze. Die blauen Teilgruppen entsprechen dabei der weiteren Einteilung im Haupttext. Die Hilfsgruppen (orange) dienen der genauen Einordnung. Teile der Abbildung wurden entnommen aus [311–314].

Interaktionen zwischen den Atomen der Schnittstelle. Dabei unterscheiden sich die Ansätze im Umfang der berücksichtigten Interaktionen und Atome, da gegebenenfalls auch sinnvoll ausgewählte Teilmengen die relevanten Aspekte repräsentieren. Die Gewichtung und Parametrisierung der einzelnen kovalenten und nicht-kovalenten Wechselwirkungen wird dabei von experimentellen Messungen oder quantenmechanischen Berechnungen abgeleitet [315].

5.1.1 Wissensbasierte Paarpotentiale

Innerhalb der Gruppe der wissensbasierten Paarpotentiale gibt es verschiedene Komplexitätsgrade der Nutzung von Informationen, die ebenfalls in die Übersicht in Abbildung 5.1 integriert sind. So nutzen die einfachen statistischen Paarpotentiale die Distanzinformation der Atom- oder Residuenpaare ausschließlich zur Unterscheidung zwischen Interaktionen und unbedeutenden Paaren. Sie gelten als distanzunabhängig, da die Distanzinformation nicht maßgebend in das eigentliche Potential eingeht. Bei den distanzbasierten Ansätzen, die nach Art und Umfang der Nutzung der atomaren Informationen unterteilt werden, fließt die Distanzinformation hingegen als wichtiger Bestandteil in das Potential ein. So verwenden die reduzierten distanzbasierten Ansätze nur einen Teil der zur Verfügung stehenden Atome oder eine reduzierte Pseudoatom- oder Residuendarstellung, während die vollständigen distanzbasierten Ansätze die Gesamtheit der zur Verfügung stehenden schweren Atome der Schnittstelle inkorporieren.

Distanzunabhängige Paarpotentiale Den distanzunabhängigen Paarpotentialen, die sich in der Literatur finden, liegen unterschiedliche aber in allen Fällen sehr kleine Datensätze von Protein-Nukleinsäure-Komplexen zugrunde. Zur Bewertung werden alle Residuenpaare der jeweils 30 bis 60 Strukturen extrahiert und entweder zur Menge der bindenden oder nicht-bindenden Paare eingeordnet. Als Grundlage für diese Einordnung werden abhängig vom Ansatz der geometrische Abstand der beteiligten Residuen, die Intensität der van-der-Waals-Interaktionen oder die Anzahl der ausgebildeten Wasserstoffbrückenbindungen genutzt. Ein direktes Einfließen dieser Eigenschaften in die Potentiale findet jedoch nicht statt. Eine Unterscheidung zwischen spezifischen Interaktionen mit den jeweiligen Nukleobasen und unspezifischen Interaktionen mit den Atomen des Nukleinsäurerückgrats wird dabei in der Regel vorgenommen. Nach Abschluss dieser Klassifizierung wird für jedes theoretisch mögliche Residuenpaar eine Präferenz abgeleitet, die deren Neigung zur Ausbildung einer Bindung beschreibt [319–325]. Die Präferenzen werden schließlich direkt oder durch einfache mathematische Nachbearbeitungsschritte in die finalen Paarpotentiale überführt. Für die Anwendung der distanzunabhängigen Paarpotentiale ist keine Distanzinformation notwendig.

Die Analysen zeigen in Übereinstimmung, dass die atomaren Kontakte zum Nukleinsäurerückgrat gegenüber den Kontakten zu den eigentlichen Nukleobasen bevorzugt werden und dass auf Seiten der Proteine die Seitenketten von größerer Bedeutung sind. Bei den Aussagen über konkrete paarweise Präferenzen und die Gewichtung von Wasserstoffbrückenbindungen und van-der-Waals-Interaktionen als Ursachen für die Bindung widersprechen sich die einzelnen Studien jedoch deutlich [324]. Diese Widersprüche zeigen stellvertretend zwei Hauptprobleme der genannten Studien auf, welche die Allgemeingültigkeit der darin ermittelten Präferenzen infrage stellen. Zum einen sind die sehr kleinen, eingesetzten Datensätze, die sich durchaus leicht überlappen können, nicht für die Generalisierung auf die gesamte Komplexklasse geeignet. Zum anderen zeugen die variierenden Parametrisierungen bei der Definition der atomaren Kontakte, wie sie in den Studien verwendet wurden, vom fehlenden Konsens über die Einflüsse der einzelnen physikochemischen Faktoren auf die Bindung. Entsprechend weisen die meisten dieser Ansätze eine Erfolgsquote auf, die nicht signifikant über dem Zufall liegt. Die wenigen Vertreter, deren Ergebnisse als statistisch signifikant bewertet wurden, sind aufgrund der schwachen Testkriterien jedoch kritisch zu beurteilen. So empfiehlt der Autor eines Ansatzes die Kombination mit der energiebasierten Bewertungsfunktion der verwendeten Dockingsoftware zum Erhalt optimaler Ergebnisse [44]. Neben den oben genannten Publikationen existieren noch weitere, die sich mit der Charakterisierung der Protein-Nukleinsäure-Bindung beschäftigen. Diese bestimmen jedoch ausschließlich Präferenzen für die Bindungsneigung der Aminosäuren [44; 326; 327] oder leiten aus den statistischen Betrachtungen keine Kennwerte ab [328–330]. In beiden Fällen können die Ansätze nicht für die Bewertung von Komplexen eingesetzt werden.

Die publizierten, distanzunabhängigen Paarpotentiale können zwar einen Teil zum Verständnis der Protein-Nukleinsäure-Bindung beitragen, die aus diesen Beiträgen gezogenen Schlüsse sollten jedoch mit äußerster Vorsicht betrachtet werden. Sie müssen kritisch beurteilt und stets mit anderen Studien in Verhältnis gesetzt werden. Trotz der Betrachtung einzelner atomarer Wechselwirkungen ist die erreichbare molekulare Auflösung der distanzunabhängigen Paarpotentiale durch die residuenweise Bestimmung der Präferenzen nicht ausreichend für die Bewertung makromolekularer Komplexstrukturen. Dazu kommen die Widersprüche zwischen den einzelnen, durchgeführten Studien, sodass sich zusammenfassend sagen lässt, dass die einfachen statistischen Paarpotentiale nicht zur Gütebewertung von Protein-Nukleinsäure-Interaktionen im praktischen Umfeld geeignet sind.

Reduzierte distanzbasierte Paarpotentiale Der Hauptunterschied der reduzierten distanzbasierten Paarpotentiale zu den eben vorgestellten einfachen Paarpotentialen ist der Einfluss der Distanzinformation auf deren Erzeugung und Anwendung. Die Häufigkeiten der einzelnen Atom- bzw. Residuenpaare werden folglich immer im Kontext ihrer geometrischen Distanzen betrachtet. Die geläufigste Herangehensweise ist dabei die Definition eines beidseitig begrenzten *CutOff*-Bereiches, in dem Interaktionen durch das Verfahren berücksichtigt werden. Durch die weitere Unterteilung dieses Bereichs in eine feste Anzahl von nicht-überschneidenden Segmenten, sogenannter Distanzschalen, wird die Berechnung gegenüber einer kontinuierlichen Berücksichtigung der Distanz mit praktisch vernachlässigbarem Verlust deutlich vereinfacht. Der Ablass von kontinuierlichen Distanzwerten ermöglicht zudem auch bei kleinen Grunddatenmengen eine ausreichende Datendichte für die darauffolgenden Berechnungen.

Unter den reduzierten distanzbasierten Konzepten befindet sich ein Vertreter, der zwar die Reduktion auf die Residuen durchführt, zusätzlich jedoch den molekularen Kontext auf Seiten der Nukleinsäure in die Berechnung einfließen lässt. Dies geschieht durch die Bildung von Nukleotid-Triplets, auf deren Basis die Ableitung der Potentiale erfolgt. Um auch einzelne Nukleotide und auftretende Paare in dieses System zu integrieren, wurde ein Pseudonukleotid eingeführt, welches unbesetzte Positionen eines Triplets belegen kann [331]. Obwohl sich bereits mehrfach gezeigt hat, dass die Unterscheidung zwischen spezifischen und unspezifischen Wechselwirkungen biologisch begründet ist [324], unterscheiden die Autoren nicht zwischen Kontakten der Nukleobasen und des Rückgrats. In der Nachbetrachtung der Studie schlagen sie jedoch selbst eine Erhöhung der molekularen Auflösung vor [331]. Wie die Tripletdarstellung zur Inkorporation der kontextuellen Information jedoch auf ein atomares Level übertragen wird, bleibt offen. Die beiden Beschreibungskonzepte *Decoys As the Reference State Potential* (DARS-RNP) und *Quasi-chemical Potential* (QUASI-RNP) nutzen eine reduzierte Darstellung der Residuen, welche für jede Aminosäure zwischen einer und drei Pseudoatomen und für jedes Nukleotid zwischen drei und vier Pseudoatomen verwendet. Durch die Verteilung dieser Pseudoatome innerhalb der Residuen wird eine Unterscheidung zwischen dem jeweiligen Rückgrat und den Seitenketten beziehungsweise Nukleobasen erreicht. In die Berechnung der Potentiale fließen neben Termen für Distanz und Winkel der Kontakte auch Informationen über den Bindungsort und das Auftreten von strukturellen Konflikten ein. Der Referenzzustand wird im Falle von QUASI-RNP über die molaren Anteile der Pseudoatomkontakte bestimmt und im Fall von DARS-RNP aus einem Ensemble von *Decoy*-Strukturen abgeleitet [332]. Das letzte Verfahren dieser Gruppe stützt sich ausschließlich auf die im Komplex vorkommenden Wasserstoffbrückenbindungen. Nach dem Einfügen der impliziten Wasserstoffatome werden die festgestellten Wasserstoffbrücken nach beteiligten Atomtypen unterschieden und durch drei Winkel sowie ihre Bindungslänge beschrieben [333].

Vollständige distanzbasierte Paarpotentiale Werden alle schweren Atome in die Berechnung einbezogen, so spricht man von vollständigen distanzbasierten Paarpotentialen. Bei dieser Herangehensweise tritt kein systematischer Informationsverlust auf, wie es bei den oben genannten Verfahren durch die Distanzunabhängigkeit und die Reduktion der Modellkomplexität der Fall ist.

In der Literatur werden drei Ansätze beschrieben, die auf Basis der *Distance-scaled, Finite, Ideal-gas Reference* (DFIRE)-Referenzfunktion [334; 335] die Information aller Atome nutzen, die zu einem von 19 definierten Atomtypen gehören. Diese Atomtypen basieren auf der Einteilung der Software SYBYL [336] und sind so gewählt worden, dass Protein-Nukleinsäure-Interaktionen

davon abgedeckt sind. Basierend auf der einfachen Anwendung des DFIRE-basierten Paarpotentials [337] wurde zunächst ein Bayesscher Ansatz entwickelt. Auch wenn dieser durch die konstante *a priori*-Verteilung einiges Potential der Bayesschen Statistik ungenutzt lässt [338], trägt er über einen Korrekturterm dem Fall geringer Vorkommen von Atompaaren Rechnung [339]. Für die serverbasierte Weblösung DDNA 2 wurden schließlich eine Reihe weiterer Korrekturterme auf die originale DFIRE-Funktion angewendet und auf ihren Effekt hin überprüft. So entstanden cFIRE durch Hinzufügen eines Korrekturglieds für selten vorkommende Atompaare und vFIRE durch eine Adaption in der Berechnung der Volumina einzelner Distanzschalen an den Spezialfall der Protein-Nukleinsäure-Komplexe, sowie die Kombination beider, vcFIRE [315]. Alternativ zu DFIRE wurde zur Erzeugung der Potentiale *Distance- and Environment-dependent, Coarse-grained and Knowledge-based RNA/Protein* (DECK-RP) ein neuer Referenzzustand entwickelt. Dieser nutzt die Informationen aus *Decoy*-Strukturen, wobei die Aminosäuren und Nukleotide entsprechend ihrer Eigenschaften und Sekundärstrukturbeteiligung zusammengefasst und ihre Schnittstellenpräferenzen einbezogen werden [340].

Im Gegensatz zu allen vorgenannten wissensbasierten Paarpotentialen fanden zwei neue Gedanken initial im Bereich der Bewertung von Protein-Protein-Interaktionen Einzug. Der erste Gedanke beschäftigt sich mit der Nutzung der Spezifität als zusätzliche Informationsquelle der Potentialerzeugung neben der ohnehin schon verwendeten Affinität. Der zweite Gedanke erweitert den Erzeugungsprozess um ein Optimierungsverfahren [341]. Beide Gedanken wurden unter dem Namen *Specificity and Affinity of the Protein-Nucleic acid Interactions* (SPA-PN) umgehend auf Protein-Nukleinsäure-Komplexe übertragen. Für die Bewertung der Spezifität ist dazu jedoch für jeden Komplex der Trainingsmenge eine große Anzahl von *Decoy*-Strukturen erforderlich [342]. In den vorangegangenen wissensbasierten Potentialen waren *Decoy*-Strukturen allenfalls als optionale Quelle zur Berechnung eines Referenzzustands dienlich [332]. Nach Berechnung der initialen, DFIRE-basierten Potentiale werden diese durch den iterativen Optimierungsprozess unter Zuhilfenahme der Spezifität schrittweise angepasst [342]. Ein ähnliches Konzept der Optimierung wurde bei der Herleitung der wissensbasierten Potentiale *Iterative Knowledge-based Scoring Function for Protein-Protein interactions* (ITScore-PP) für Protein-Protein-Interaktionen angewendet [343]. Die Übertragung dieser Potentiale in die Domäne der Protein-Nukleinsäure-Komplexe mit dem Namen *Iterative Knowledge-based Scoring Function for Protein-RNA interactions* (ITScore-PR) beschränkte sich jedoch auf RNA als Bindepartner. Die ITScore-Methodik greift auf Ansätze der statistischen Mechanik zurück, um die initialen Potentiale abzuleiten und die Korrekturterme für die Optimierung bereitzustellen. Sie umgeht damit das Problem, einen Referenzzustand definieren zu müssen. Als Grundlage für den Optimierungsprozess wird ebenfalls eine große Bibliothek von *Decoy*-Strukturen für jeden Eingabekomplex gebraucht [344].

5.1.2 Molekularmechanische Bewertung

Im Gegensatz zu den wissensbasierten Paarpotentialen gründet sich die molekularmechanische Bewertung nicht auf eine statistische Analyse bekannter Strukturen. Zwar werden auch bekannte Strukturen für die eingesetzte semi-empirische Ableitung der Bewertungsmaße benötigt, jedoch sind dies nicht nur Proteine und Nukleinsäuren, sondern auch andere, kleine Moleküle, deren Eigenschaften zum Abbilden grundlegender Interaktionen geeignet sind. Das Ableiten der nötigen Kennwerte erfolgt durch die Kombination verschiedener physikalischer Interaktionsterme mit dem Anspruch einer absoluten Bewertung auf energetischer Basis. Die Gewichtung und Parametrisierung dieser Terme wird dabei durch zahlreiche experimentelle

Messungen und quantenmechanische Berechnungen bestimmt [315]. Auch wenn die molekularmechanischen Bewertungsmodelle in der Regel molekulare Zielgruppen für ihre Anwendung definieren, erfordern sie ihrer Bestimmung und ihres Ursprungs gemäß ein weitaus größeres Maß an Allgemeingültigkeit als die bekannten wissensbasierten Ansätze.

Dem primären Einsatz dieser Ansätze in der molekulardynamischen Simulation sind einige weitere Anforderungen geschuldet. So muss die Solvatisierung molekularer Strukturen korrekt in der Bewertung berücksichtigt werden. Die Approximation der zeitabhängigen Schrödinger-Gleichung bei der Bestimmung der wirkenden Kräfte innerhalb des Simulationssystems erfordert eine sehr hohe Genauigkeit und Verlässlichkeit der molekularmechanischen Bewertung [345]. Die Höhe dieser Anforderungen kann von einem wissensbasierten Paarpotential und den davon abgeleiteten Pseudoenergien nicht erfüllt werden. Die molekularmechanischen Bewertungsmaße, die im Bereich der molekulardynamischen Simulation zumeist Kraftfelder genannt werden, setzten sich dabei aus Interaktionstermen zusammen, die neben Elektrostatik und van-der-Waals-Interaktionen Bindungslängen und -winkel sowie Torsionswinkel einbeziehen. Sie sind dabei primär zur Bewertung von Einzelmolekülen und Molekülsystemen geeignet. Um ausschließlich die Bindung der Komplexpartner zu beschreiben, wie es hier erforderlich ist, wird die Differenz der bestimmten Energien aus gebundenem Molekülkomplex und ungebundenen Einzelkomponenten gebildet.

Neben vielen weiteren befinden sich unter den Kraftfeldern mit Eignung für makromolekulare Komplexe die bekannten Vertreter *Assisted Model Building with Energy Refinement* (AMBER) [346], *Groningen Molecular Simulation* (GROMOS) [347], *Chemistry at Harvard Macromolecular Mechanics* (CHARMM) [348] und *Optimized Potentials for Liquid Simulations* (OPLS) [349]. Neben den reinen Protein-Parametrisierungen existieren für diese Kraftfelder auch Parametersätze für den gemischten Einsatz auf Protein- und Nukleinsäurebestandteile [350]. Die Kraftfelder unterliegen einer anhaltenden Entwicklung. In der Regel wurden neue Erkenntnisse in mehreren Revisionen in die Kraftfelder eingepflegt, sodass die auftretenden Fehler verringert und zusätzliche Molekülgeometrien unterstützt wurden. Die zugehörigen Referenzimplementierungen sind jeweils in molekulare Simulationssysteme einbettet. Auch wenn der Fokus hier auf der komplexen Molekulardynamik liegt, bieten diese Softwarepakete die Möglichkeit, Energieberechnungen und -minimierungen separat durchzuführen. Die Nutzung der Kraftfelder bleibt nicht auf die jeweilige Referenzimplementierung beschränkt, die zum Teil schon mehrere Kraftfelder unterstützt [351], sondern ist auch anderen Bereichen der molekularen Simulation zugänglich. Beispiel hierfür ist das informationsgetriebene Dockingsystem *High Ambiguity Driven Biomolecular Docking* (HADDOCK). Es verwendet für die interne Bewertung der generierten *Decoy*-Strukturen während eines Dockinglaufes ein Bewertungsschema, welches auf dem Kraftfeld OPLS aufbaut [352]. Die endgültige Auswahl eines besten Kraftfeldes ist dabei jedoch kaum möglich, da keine Form der exakten Herleitung genutzt wird, sondern unterschiedliche Aspekte entsprechend gewichtet in die theoretisch-physikalische Herleitung einfließen.

5.1.3 Auswahl der Konzepte für die weitere Betrachtung

Für den bevorstehenden Vergleich musste nun eine geeignete Auswahl aus den zahlreichen vorgestellten Konzepten gefunden werden. Dazu wurden unter den wissensbasierten Ansätzen ausschließlich die vollständigen, distanzbasierten Verfahren in die nähere Wahl genommen, da sowohl bei den distanzunabhängigen als auch bei den reduzierten Strategien zu viele Informationen der zu bewertenden Molekülkomplexe nicht in die Bewertung einfließen. Unter den vollständigen, distanzbasierten Verfahren fanden sich neben der Vielzahl teils schon älterer, ein-

facher Modelle zwei Ansätze, die durch ein angeschlossenes Optimierungsverfahren eine hohe informationelle Ausbeute und damit eine zielführende Bewertung versprachen. Diese beiden Verfahren, SPA-PN und ITScore-PR, zielen auf unterschiedliche Optimalitätskriterien hin und bildeten daher sinnvolle Vergleichspartner.

Die molekularmechanischen Ansätze haben sowohl eine überschneidende Datenbasis als auch ein in weiten Teilen identisches Grundkonzept. Die aus dieser Konstellation erwachsenden Unterschiede sind zwar für den Ausgang einer sensiblen molekulardynamischen Simulation von entscheidender Wichtigkeit [353], sollten jedoch bei einer reinen energetischen Bewertung von Komplexstrukturen nur geringfügige Unterschiede zu Tage bringen. Selbiges ist auch für die unterschiedlichen Revisionen der einzelnen Kraftfelder anzunehmen. Für das am häufigsten eingesetzte [354] und damit am besten erprobte Kraftfeld AMBER konnte in der Molekulardynamik bereits in seinen frühen Revisionen ein sehr konsistentes Verhalten, sowohl bei Nukleinsäuren als auch bei Proteinen, gezeigt werden [355]. Es wurde daher Teil des Vergleichs. Auch wenn bei der Verwendung des OPLS-Kraftfeldes bei Nukleinsäuren eine gewisse Vorsicht empfohlen wird [351], wurde der auf OPLS basierende HADDOCK-Score ebenfalls in den Vergleich einbezogen [352]. Hier war jedoch weniger die Abdeckung eines weiteren Kraftfeldes sondern eher der ohnehin geplante Einsatz des Dockingsystems HADDOCK der Grund.

5.2 Vorstellung und Herleitung der ausgewählten Bewertungsmodelle

5.2.1 Die SPA-PN-Potentiale

Zur Ausübung einer biologischer Funktion durch Prozesse der molekularen Erkennung sind sowohl Stabilität als auch Spezifität der Bindung im Komplex wichtig. Wie bereits in Abschnitt 5.1 besprochen, fließt in die meisten Bewertungsmodelle für Protein-Nukleinsäure-Bindungen lediglich die Affinität als Maßgröße der Stabilität ein, die Spezifität bleibt jedoch unberücksichtigt. Die SPA-PN-Potentiale bieten über die Einbindung der Spezifität daher eine wichtige Ergänzung der bisherigen Verfahrenslandschaft. Die Herleitung des Paarpotentials SPA-PN wird zwar in der dazugehörigen Publikation [342] beschrieben, jedoch finden sich weder im Hauptdokument noch in den ergänzenden Materialien die finalen Werte des Potentials oder Hinweise auf deren Verbleib. Um diese zu erhalten, wurde das Verfahren anhand der publizierten Beschreibung nachvollzogen.

Intrinsische Spezifität Der Hauptgrund für die alleinige Fokussierung vieler Ansätze auf die Affinität liegt in der Natur der konventionellen Spezifität. Sie gibt den Grad der Favorisierung einer bestimmten molekularen Bindung gegenüber kompetitierenden Bindungspartnern an. Zwar ist sie damit relativ leicht definiert, ihre Quantifikation birgt jedoch einige Probleme. Da in der Regel nur experimentelle Daten über die natürlich vorkommenden Komplexe vorliegen, nicht jedoch über die alternativen Bindungsmöglichkeiten, ist die Berechnung der konventionellen Spezifität im Sinne der wissensbasierten Paarpotentiale praktisch nicht möglich [341]. Aus diesem Grund entstand ein neues Konzept für die intrinsische Charakterisierung der Spezifität. Diese beschreibt jeweils für einen molekularen Komplex den Grad der Favorisierung der natürlichen Konformation gegenüber alternativen, unnatürlichen Konformationen derselben Bindepartner [356; 357]. Sie verringert damit die zu nehmende Hürde, da alternative Bindungspartner für ihre Berechnung nicht mehr notwendig sind. Es konnte gezeigt werden, dass zwischen intrin-

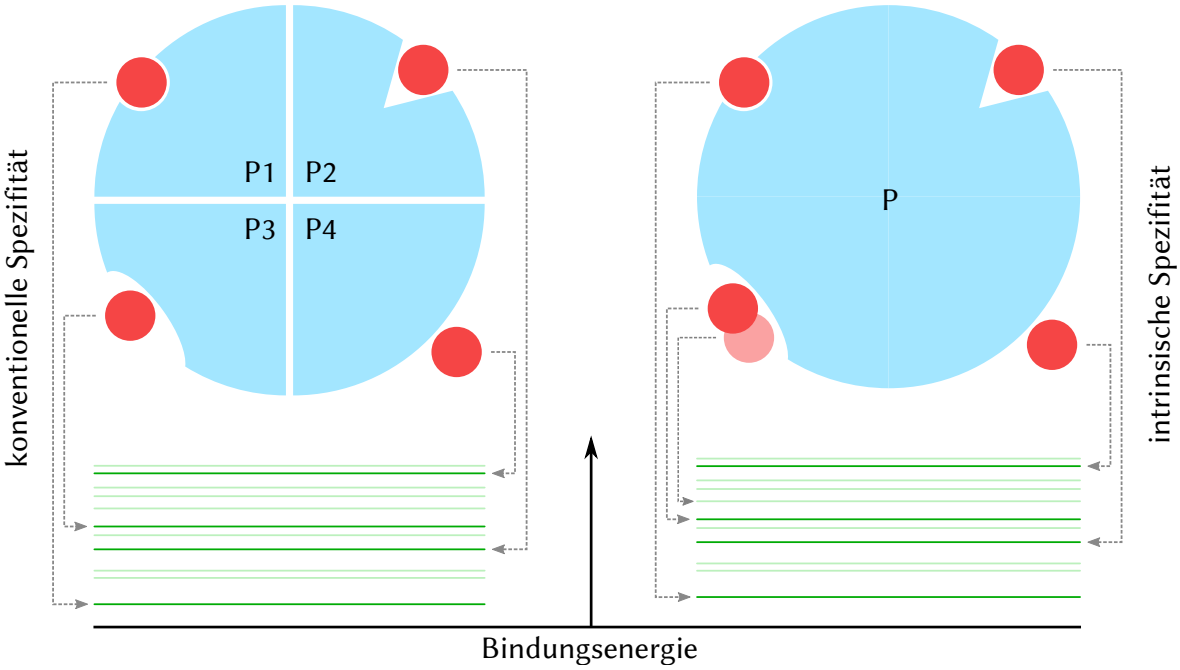


Abb. 5.2: Vergleichende Illustration des Übergangs von der konventionellen (links) zur intrinsischen (rechts) Spezifität. Bei der Bindung eines Liganden mit verschiedenen Proteinrezeptoren (links) zeigt sich die konventionelle Spezifität in den Unterschieden der Bindungsaffinitäten an den jeweiligen Rezeptoren (P1 bis P4). Bei der Bindung desselben Liganden an verschiedenen Bindeflächen eines hinreichend großen Proteinrezeptors (rechts, P) zeigt sich die intrinsische Spezifität in den Unterschieden der Bindungsaffinitäten zwischen dem originalen und den anderen beobachtbaren Bindungsmodi. Die Bindungsmodi umfassen dabei sowohl unterschiedliche Lokalisationen der Bindung auf dem Rezeptor als auch verschiedene räumliche Ausrichtungen des Liganden. Die Bindungsaffinitäten sind im unteren Teil der Abbildung als vereinfachtes Energiespektrum (grün) dargestellt. Die Abbildung ist gestaltet nach [342; 359].

sischer und konventioneller Definition der Spezifität eine Korrelation besteht [358], die auf die gute Anwendbarkeit des neuen Konzepts schließen lässt. Zum Verständnis bietet Abbildung 5.2 eine vergleichende, schematische Übersicht über die beiden genannten Formen der Spezifität.

Es ist bekannt, dass beim Bindungsprozess ähnlich wie beim Faltungsprozess eine raue, trichterförmige Energielandschaft durchlaufen wird, an deren Ende der native Zustand mit einer sehr geringen freien Energie liegt [356]. So ist zu beobachten, dass die native Komplexstruktur eine geringere freie Bindungsenergie ausweist als nahe- und nicht-native Strukturen. Die entstehende Verteilung der freien Energien ähnelt in einem solchen Fall der gaußschen Normalverteilung. Auf dieser Basis ist der *Intrinsic Specificity Ratio* (ISR) in der Lage, den Grad der intrinsischen Spezifität nach Formel 5.1 aus der Verteilung zu bestimmen. Dabei bezeichnet δE die durchschnittliche Energielücke zwischen nativer Konformation und den nahe- und nicht-nativen Konformationen des *Decoy*-Strukturensembles, während ΔE anhand der Breite der Energieverteilung im Ensemble eine Aussage bezüglich ihrer Rauheit macht. S ist schließlich die konformationelle Entropie. Basierend auf dieser Definition kann die intrinsische Spezifität berechnet werden, sobald eine hinreichend große Menge nahe- und nicht-nativer Strukturen vorhanden ist [342].

$$\mathrm{ISR} = \frac{\delta E}{\Delta E \sqrt{S}} \tag{5.1}$$

Tab. 5.1: Auflistung der relevanten Atomtypen pro makromolekularem Typ in der Systematisierung von SYBYL [336]. Nicht alle Kombinationen der hier gegebenen Atomtypen wurden als relevant eingestuft.

Makromolekül	Beteiligte Atomtypen
Nukleinsäure	C_3, C_{ar}, N_{ar}, N_{pl3}, O_2, O_3, P_3
Protein	C_2, C_3, C_{ar}, C_{cat}, N_3, N_4, N_{am}, N_{ar}, N_{pl3}, O_2, O_3, O_{co2}, S_3

Datengrundlage Zum Aufbau eines Trainings- und Testdatensatzes wurden alle Protein-Nukleinsäure-Komplexe der *Nucleic Acids-Protein Interaction Database* (NPIDB) [360] nach strengen Qualitätskriterien gefiltert. Der Empfehlung der Originalautoren folgend wurden nur Strukturen erhalten, welche mittels *X-Ray Diffraction* (XRD) mit einer hohen Auflösung aufgeklärt wurden. Die Grenzwerte wurden dabei jedoch mit $2{,}0\,\text{Å}$ (DNA) beziehungsweise $2{,}5\,\text{Å}$ (RNA) restriktiver gewählt als vorgeschlagen [342]. Durch die stets wachsende Datenbasis der NPIDB war diese erhöhte Anforderung durchsetzbar ohne damit die Größe des selektierten Datenbestandes maßgeblich zu verringern. Ferner wurden Komplexe mit über sechs makromolekularen Ketten aus der Betrachtung ausgeschlossen, da ungewollte Wechselwirkungen zwischen den Ketten ansonsten zunehmend als Störgröße in der Berechnung auftreten würden. Nach dem Bezug der Strukturdateien aus der PDB [305] wurden im Rahmen einer ersten Bereinigung irrelevante Gruppen wie Wassermoleküle, einzelne Metallionen sowie kleine und nichtorganische Nichtstandardreste entfernt. Verbleibende organische Nichtstandardreste in den makromolekularen Ketten konnten vom Verfahren nicht verarbeitet werden, sodass die entsprechenden Ketten entfernt wurden. Aufgrund zu kleiner bindungsfähiger Oberflächen wurden schließlich diejenigen Ketten aus den Strukturen verworfen, die weniger als 100 schwere Atome besaßen.

Nach dem Einlesen der Strukturen durch eine über *Java Native Interface* (JNI) angebundene Version der Software OpenBabel [361] wurden die paarweisen Häufigkeiten der Atomtypen distanzunabhängig, jedoch innerhalb der äußeren Grenzen bestimmt. Durch den empfohlenen Schwellwert von mindestens 600 Vorkommen [342] wurden relevante Atomtyppaare ausgewählt. Am Aufbau der insgesamt 84 Paare waren die Atomtypen aus Tabelle 5.1 beteiligt. Nun wurden die Abstandsbereiche mit einer Schalengröße von $\Delta r = 0{,}3\,\text{Å}$ und einer Ausdehnung von $2{,}2\,\text{Å}$ bis $R = 8{,}2\,\text{Å}$ entsprechend der Originalpublikation festgelegt. Jede der 20 resultierenden Distanzschalen wird im folgenden mit ihrem inneren Radius r angesprochen und erstreckt sich entsprechend der Schalenbreite bis $r + \Delta r$ [342]. Für jedes Paar von Protein- und Nukleinsäureketten wurde schließlich mittels eines Schwellwerts von 500 atomaren Kontakten innerhalb der definierten Schalen überprüft, ob es sich um ein hinreichend interagierendes Kettenpaar handelte. Interagierende Ketten wurden extrahiert und im weiteren als Datengrundlage verwendet. Aus der Grunddatenmenge wurden 1000 Strukturen als Trainingsmenge und weitere 100 als Testdatensatz bereit behalten.

Erzeugen von Decoy-Strukturen Nachdem diese Datengrundlage geschaffen war, wurden durch Anwendung von Standardprotokollen der Softwaresuite Rosetta [362] *Decoy*-Strukturen für alle Komplexe erzeugt [342]. Beim Nachvollziehen der Originalpublikation eröffnete sich jedoch, dass die Docking- und Bewertungsfunktionalität der *Rosetta Suite* keine Protein-Nukleinsäure-Komplexe unterstützte. Eine Modifikation oder der Einsatz zusätzlicher Software wurde jedoch von den Autoren nicht beschrieben [342]. Bei der Inspektion des Quellcodes der *Rosetta Suite* stellte sich heraus, dass die zum Durchführen des Dockings notwendige Bewertungsfunktion Nukleinsäuren nicht bewerten konnte, da die Definition des Nukleinsäurerückgrats fehlte.

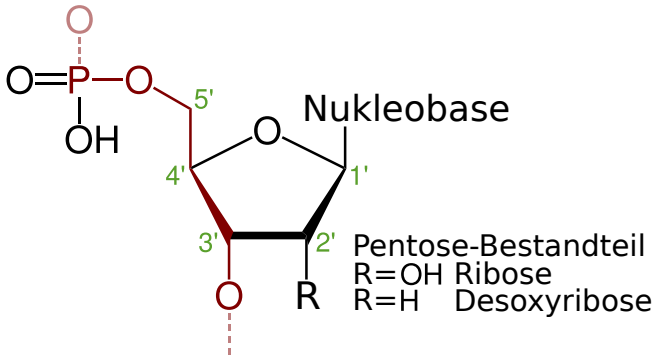
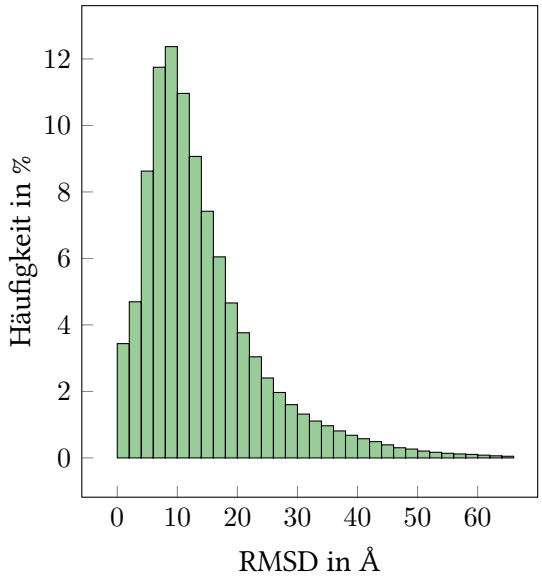


Abb. 5.3: Schematische Darstellung eines Nukleotids mit Nummerierung der Kohlenstoffatome (grün) und Hervorhebung des Nukleinsäurerückgrats (rot). Anteile der anschließenden Nukleotide sind durch Verringerung der Farbintensität kenntlich gemacht.

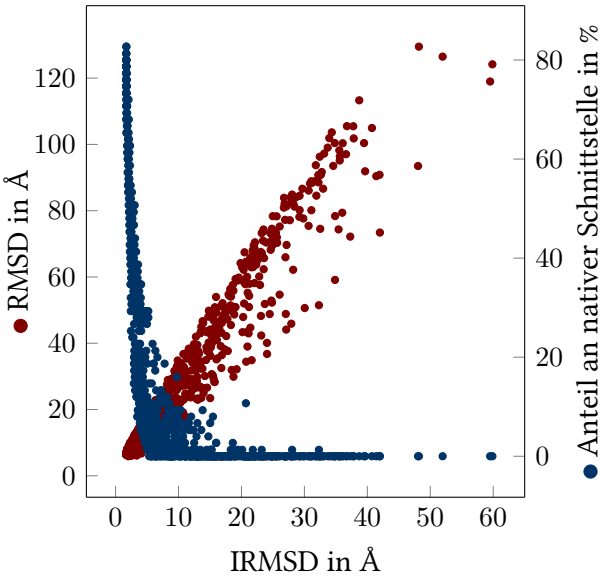
Entsprechend wurde der Dockingprozess automatisch abgebrochen. Die fehlende Definition des Nukleinsäurerückgrats wurde durch die Atome O3', C3', C4', C5', O5', und P entsprechend Abbildung 5.3 nachträglich festgelegt und durch eine eigene Implementierung in Rosetta integriert.

Die Komplexstrukturen konnten nun den notwendigen Vorbereitungsschritten unterzogen werden, welche neben dem *Prepack*- und *FastRelax*-Protokoll, auch die separat ausführbare Verfeinerungsphase des eigentlichen Dockingprotokolls umfassten. So vorbereitet wurden zu jedem Protein-Nukleinsäure-Komplex der Trainings- und Testmenge, parallelisiert durch das *Java Parallel Processing Framework* (JPPF), 1000 *Decoy*-Strukturen erzeugt. Um dabei hinreichend viele nahe-native Konformationen zu erhalten, wurden die Grenzen der initialen, zufälligen Rotation und Translation mit $3°$ beziehungsweise 4 Å pro Koordinatenachse sehr eng gewählt. Da diese Werte von der *Rosetta Suite* eher wie ein Richtwert als einen wirklichen Grenzwert umgesetzt werden, streuten die tatsächlichen Werte von Rotation und Translation um diese Vorgaben. Dabei kam es zum Teil zu erheblichen Überschreitungen, welche für die Erzeugung zahlreicher nicht-nativer *Decoy*-Strukturen verantwortlich waren. Die Verteilung der strukturellen Abweichungen des gesamten *Decoy*-Ensembles in Abbildung 5.4a verdeutlicht diesen Sachverhalt. Der modifizierte Komplex wurde schließlich in einem 6-zyklischen Dockinglauf mit anschließender Minimierung wieder verbunden. Ein zweites Dockingprotokoll lieferte die vom Optimierungsalgorithmus verwendeten Bewertungsmaße. Die Ansteuerung der unterschiedlichen Werkzeuge der *Rosetta Suite* erfolgte zentral durch *Extensible Markup Language* (XML)-Scripte über die *RosettsScripts Engine* [363]. Die erreichte Mischung aus vielen nahe-nativen und einigen deutlich unterschiedlichen Strukturen (siehe Abbildung 5.4a) ist für die weitere Optimierung gut geeignet. Eine Inspektion der weiteren Eigenschaften der *Decoy*-Strukturen in Abbildung 5.4b zeigt keine Auffälligkeiten. Sowohl die Korrelation zwischen den Abweichungen der Gesamtstrukturen und zugehörigen Regionen der Schnittstellen als auch die Relation zum Konservierungsgrad der Schnittstelle zeigen ein erwartetes Verhalten.

Ableitung der initialen, beobachteten Potentiale Basierend auf der Einteilung in 20 Distanzschalen und 84 relevante Atomtyppaare wurden ohne Einbezug der Spezifität zunächst die 1680 beobachteten Potentiale u_k^{obs} berechnet. Dazu wurden in allen nativen Strukturen $m \in M$ die absoluten Häufigkeiten der Kombinationen aus Atomtyppaar k und Distanzschale r als $n_k^m(r)$ bestimmt. Die Summe $N_k^m = \sum_r n_k^m(r)$ beschreibt dabei die absoluten Häufigkeiten ohne Berücksichtigung der Distanzinformation. Unter Zuhilfenahme der Volumina der Referenz-



(a) Häufigkeitsverteilung über die strukturellen Abweichungen (RMSD) der *Decoy*-Strukturen von der nativen Konformation.



(b) Zusammenhang zwischen Gesamtabweichung (RMSD, rot), Abweichung (IRMSD) und Konservierungsgrad (blau) der Kontakte der nativen Schnittstelle.

Abb. 5.4: Übersicht über die erzeugten *Decoy*-Strukturen. Die Häufigkeitsverteilung der RMSD (a, grün) bezieht sich auf alle generierten Strukturen, während die beiden Verhältnisdarstellungen in b nur diejenigen des beispielhaften Komplexes 3LWV_AD berücksichtigen. Die Diagramme zeigen ein erwartetes Verhalten und unterlegen damit den korrekten Ablauf des Erzeugungsprozesses.

schalen $V(r)$ und Referenzsphäre $V(R)$ aus Formel 5.2 wurden aus den absoluten Häufigkeiten jeweils für die Distanzschalen und die Gesamtsphäre Teilchendichten $f_k^{\text{obs}}(r)$ und $f_k^{\text{obs}}(R)$ abgeleitet. Die genaue Ableitungsvorschrift befindet sich in Formel 5.3. Die aus den Teilchendichten berechnete Verteilungsfunktion der Atompaare $g_k^{\text{obs}}(r)$ wurde schließlich entsprechend Formel 5.4 nach dem inversen Boltzmann-Prinzip [364] in die beobachteten Atompaarpotentiale $u_k^{\text{obs}}(r)$ überführt [342]. Der konstante Faktor k_BT konnte bei der Berechnung der Potentiale vernachlässigt werden, da er keinen Einfluss auf die relativen Verhältnisse der Potentiale ausübt, die im weiteren betrachtet werden. Für das Einlesen der Eingabedateien und die Berechnung der initialen Potentiale wurden eine eigene Java-Implementierungen verwendet.

$$V(r) = \frac{4}{3}\pi\left((r+\Delta r)^3 - r^3\right) \qquad V(R) = \frac{4}{3}\pi R^3 \tag{5.2}$$

$$f_k^{\text{obs}}(r) = \frac{1}{M}\sum_{m\in M}\frac{n_k^m(r)}{V(r)} \qquad f_k^{\text{obs}}(R) = \frac{1}{M}\sum_{m\in M}\frac{N_k^m}{V(R)} \tag{5.3}$$

$$g_k^{\text{obs}}(r) = \frac{f_k^{\text{obs}}(r)}{f_k^{\text{obs}}(R)} \qquad u_k^{\text{obs}}(r) = -k_BT\cdot\ln g_k^{\text{obs}}(r) \tag{5.4}$$

Ableitung der erwarteten Potentiale Bei der Ableitung der erwarteten Potentiale u_k^{exp} wurde die Annahme vorausgesetzt, dass die Menge der generierten *Decoy*-Strukturen eine Boltzmann-Verteilung um die native Struktur bildet. Die native Struktur wird dabei jeweils als Teil des Ensembles verstanden. Zur Berechnung der erwarteten Teilchendichten $f_k^{\text{exp}}(r)$ und $f_k^{\text{exp}}(R)$ wurden nun die Informationen des gesamten Ensembles eingebracht. Dabei wurde wie in Formel 5.5 dargestellt ein Gewichtungsschema auf Basis der Boltzmann-Verteilung eingesetzt. Der konstante Faktor β nahm in der Berechnung die Funktion einer inversen Boltzmann-

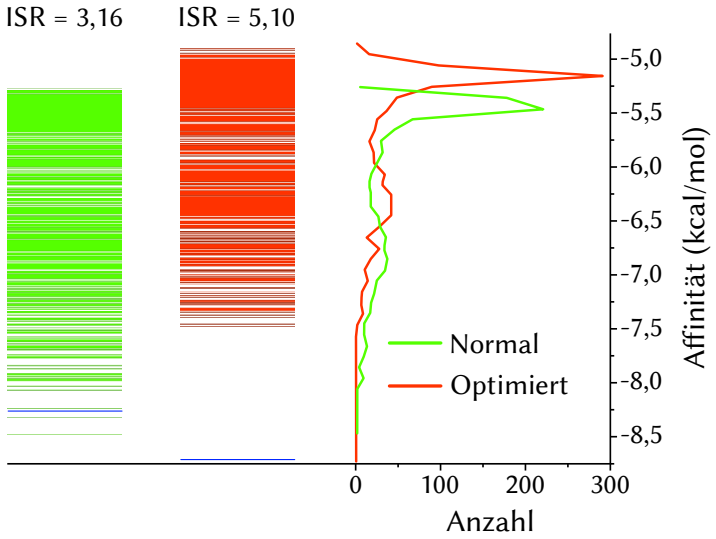


Abb. 5.5: Verteilung der berechneten Energien eines Beispielkomplexes und der zugehörigen *Decoy*-Strukturen, jeweils vor der Optimierung (grün) und danach (orange). Die native Struktur ist blau im Spektrum (links) hervorgehoben, ihr ISR ist über den Spektren angegeben. Die Optimierung führt zu einer deutlich besseren Separierung nahe- von nicht-nativen Komplexstrukturen [342].

Temperatur ein, während der Term $U_{md} = \gamma E_{md} + \lambda_{md}$ ein zusammengesetztes Potential zur Unterscheidung zwischen nativer und nicht-nativer Konformation darstellte. Es setzte sich zusammen aus der Energiebewertung E_{md}, die mithilfe der aktuellen Potentiale auf die Affinität schloss, und der über den ISR λ_{md} unter Einbezug des gesamten Strukturensembles bestimmten Spezifität. Der Faktor γ diente der Gewichtung der beiden Teilaspekte. Die Verteilungsfunktion und die eigentlichen erwarteten Paarpotentiale wurden schließlich entsprechend Formel 5.6 berechnet [342].

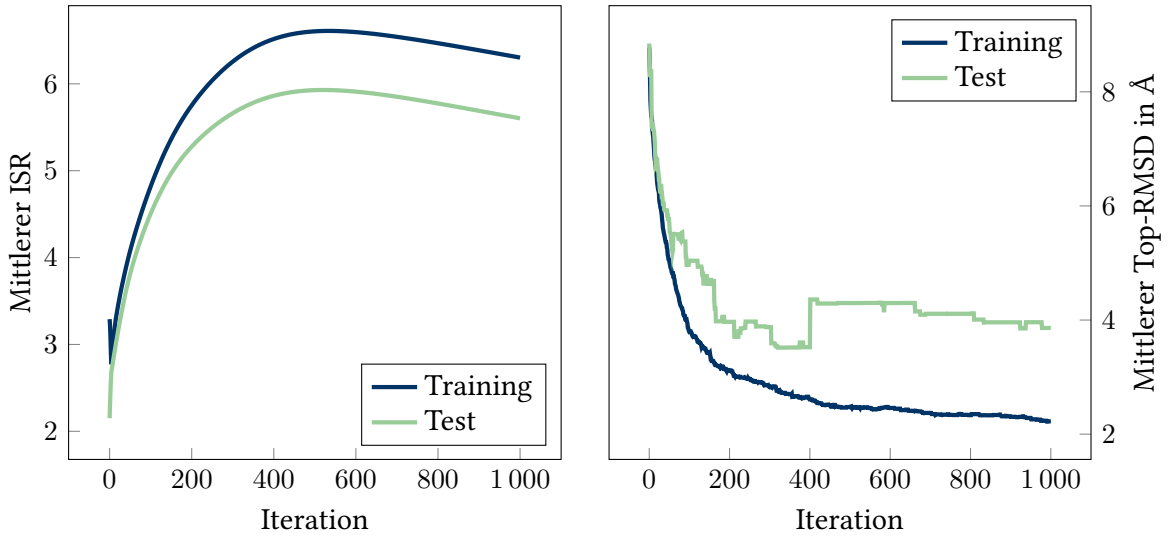
$$f_k^{\exp}(r) = \frac{1}{MD} \sum_{m \in M} \sum_{d \in D} \frac{n_k^{md}(r) \mathrm{e}^{-\beta U_{md}}}{V(r)} \qquad f_k^{\exp}(R) = \frac{1}{MD} \sum_{m \in M} \sum_{d \in D} \frac{N_k^{md} \mathrm{e}^{-\beta U_{md}}}{V(R)} \tag{5.5}$$

$$g_k^{\exp}(r) = \frac{f_k^{\exp}(r)}{f_k^{\exp}(R)} \qquad u_k^{\exp}(r) = -k_B T \cdot \ln g_k^{\exp}(r) \tag{5.6}$$

Iterative Optimierung Die finalen Potentiale SPA-PN wurden durch einen iterativen Prozess schrittweise optimiert. Dazu wurde in jeder Iteration i wie in Formel 5.7 gezeigt die Differenz $\Delta^i u_k(r)$ zwischen den erwarteten Potentialen der aktuellen Iteration und den beobachteten Potentialen bestimmt. Diese Differenz wurde schließlich nach Formel 5.8 zur Korrektur der erwarteten Potentiale genutzt, wobei die Konstante χ die Geschwindigkeit der Konvergenz regelte. Da die erwarteten Potentiale in die Berechnung des zusammengesetzten Potentials U_{md} einflossen, entstand im Optimierungssystem eine Rückkopplung, die für die korrekte Ausführung der Optimierung notwendig ist [342]. Der Effekt der Optimierung auf die ISR-Bewertung wird in Abbildung 5.5 exemplarisch gezeigt.

$$\Delta^i u_k(r) = u_k^{\mathrm{obs}}(r) - {}^i u_k^{\exp}(r) \tag{5.7}$$

$${}^{(i+1)} u_k^{\exp}(r) = {}^i u_k^{\exp}(r) + \chi \cdot \Delta^i u_k(r) \tag{5.8}$$



Mittlerer ISR **(a)** **(b)** Mittlerer Top-RMSD

Abb. 5.6: Übersicht über den Verlauf der Optimierung. Gezeigt ist die Entwicklung der Potentiale anhand zweier Beobachtungsgrößen über 1000 Iterationen. Während der ISR im Verlaufe der Optimierung bis zu einem bestimmten Punkt steigt, nimmt der Top-RMSD ab.

Der Optimierungsalgorithmus wurde freundlicherweise durch den Autor in der Programmiersprache C++ zur Verfügung gestellt [365], sodass eine Portierung in die Sprache Java durchgeführt werden konnte. Wie in Abbildung 5.6 dargestellt ist, ging die Optimierung im Mittel tatsächlich mit einer Verbesserung der Unterscheidungsleistung der Potentiale einher. Am Wachstum des mittleren ISR ist in Abbildung 5.6a zu sehen, dass durch die Potentiale die native Komplexstruktur zunehmend stärker gegenüber den *Decoy*-Konformationen präferiert wurde. Ab etwa der Iteration 500 war eine Stagnation und der beginnende Verlust dieser Fähigkeit zu verzeichnen. Offensichtlich wies das Optimierungsverfahren ein nicht-optimales Konvergenzverhalten auf, sodass zum Erhalt der optimalen Lösung ein geeignetes Abbruchkriterium erforderlich war. Das Verhalten im Testdatensatz war analog, jedoch mit etwas schwächer ausgeprägtem ISR. Ein weiteres Bewertungskriterium stellten die mittleren Top-RMSD der Strukturen dar. Für jede Eingangsstruktur wurde die auf Basis der aktuellen Potentiale am besten bewertete *Decoy*-Struktur ausgewählt und mit der nativen Struktur verglichen. Der Mittelwert dieser Abweichungen wurde in Abbildung 5.6b für jede Iteration aufgetragen. Aus der Abbildung geht hervor, dass diese mittleren Top-RMSD über die gesamte Optimierungsdauer beständig sanken. Beim Testdatensatz wurde das absteigende Verhalten im mittleren Top-RMSD durch einen Anstieg bei Iteration 400 unterbrochen. Es ist anzunehmen, dass es an dieser Stelle im Zuge der Optimierung zu einer ungünstigen strukturellen Favorisierung der Potentialbewertung gekommen ist. Nach dem Anstieg setzte sich jedoch erneut ein leicht abnehmender Trend durch.

Unter Einbezug der Kenntnis über das Vorhandensein nahe-nativer *Decoy*-Konformationen kann die rückläufige Tendenz des ISR ab Iteration 500 dadurch erklärt werden, dass deren Bewertungen durch die Potentiale denen der nativen Struktur sehr ähnlich wurden. Dies wirkte sich zwar negativ auf den ISR aus, war jedoch ein tolerierbares Verhalten im Sinne der Optimierung. Im Vergleich zwischen den initial beobachteten und den final optimierten Potentialen zeigte eine manuelle Inspektion, dass für die meisten Trainingsstrukturen eine Verbesserung des ISR

erreicht wurde. Nur ein kleiner Bruchteil wies eine leichte Verringerung des ISR auf. Ähnlich verhielt es sich auch beim Top-RMSD. Ein gepaarter, einseitiger t-Test zeigte eine signifikante Verbesserung sowohl bei den ISR ($p < 0.002$) als auch bei den Top-RMSD ($p < 0.001$).

5.2.2 Die modifizierten SPA-PN Potentiale

Nach der Implementierung des gesamten Erzeugungsprozesses wurde der Einfluss einiger Parameter auf das Verfahren untersucht, um eine Verbesserung der Beschreibung zu erwirken. Zur Modifikation kamen die Parametrisierung der Distanzeinteilung sowie die Art der Repräsentation von Atomtypen infrage.

Nutzung des molekularen Kontextes Im Bereich der wissensbasierten Potentiale wurde bereits gezeigt, dass Kontextinformationen nutzbringend in die Berechnung eingebracht werden können [331]. Neben den Atomtypen und deren Distanzen fließen in die Berechnung der SPA-PN jedoch keine kontextbezogenen Informationen ein. Ziel der Modifikation war daher die Einführung eines geeigneten molekularen Kontextes. Eine Lösung über das direkte Einbeziehen der umliegenden Atome schied aus, da die nötige Datendichte zur Ableitung der Potentiale durch die verbundene breite Streckung der Daten nicht mehr gegeben war. Über die Zuordnung der einzelnen Atome zu Residuen kann ebenfalls auf deren Kontext geschlossen werden. Da dieser jedoch über den atomaren Aufbau der Residuen schematisch stark fixiert vorlag, erfolgte bei dieser Erweiterung nur eine geringe Streckung der Daten.

Auf der Seite der Nukleinsäuren wurden dabei die Nukleobasen Thymin und Uracil zusammengefasst, da sich diese nur durch die Existenz einer Methylgruppe unterscheiden, ansonsten jedoch beide zur Gruppe der Pyrimidine gehören und die gleiche Watson-Crick-Paarungscharakteristik mit Adenin aufweisen. Die bestehende Empfehlung, zwischen Kontakten zu den Atomen der einzelnen Nukleobasen und denen zum Zucker-Phosphat-Rückgrat zu unterscheiden [324], wurde durch eine Pseudogruppe umgesetzt. Für die Nukleinsäuren existierten daher fünf Gruppen zur Erfassung des atomaren Kontextes. Auf der Seite der Proteine wird die hohe Relevanz der Seitenketten für die Bindungscharakteristik betont. Der geringen Bedeutung der Kontakte zum Rückgrat geschuldet, wurde eine gesonderte Behandlung dieser nicht durchgeführt [324]. In der sogenannten vollständige Gruppierung wurden daher genau die 20 Residuen der Proteine unterschieden. Die physikochemischen Eigenschaften der Aminosäuren boten über diese einfache Einteilung hinaus auch weitere, eigenschaftsbezogene Möglichkeiten der Gruppierung. Da die elektrostatischen Wechselwirkungen in Protein-Nukleinsäure-Komplexen von Bedeutung sind, wurde daher ebenfalls die elektrostatische Gruppierung auf ihre Eignung untersucht. Alanin, Valin, Methionin, Leucin, Isoleucin, Prolin, Tryptophan und Phenylalanin wurden dabei als unpolare hydrophobe Aminosäuren zusammengefasst, während Tyrosin, Threonin, Glutamin, Glycin, Serin, Cystein und Asparigin eine Gruppe von polaren und neutralen Aminosäuren bildeten. Schließlich wurde unter den geladenen Aminosäuren unterschieden zwischen den Säuren Glutaminsäure und Aspariginsäure, sowie den Basen Lysin, Arginin und Histidin. Ergebnis dieser Einteilung sind vier elektrostatische Gruppen.

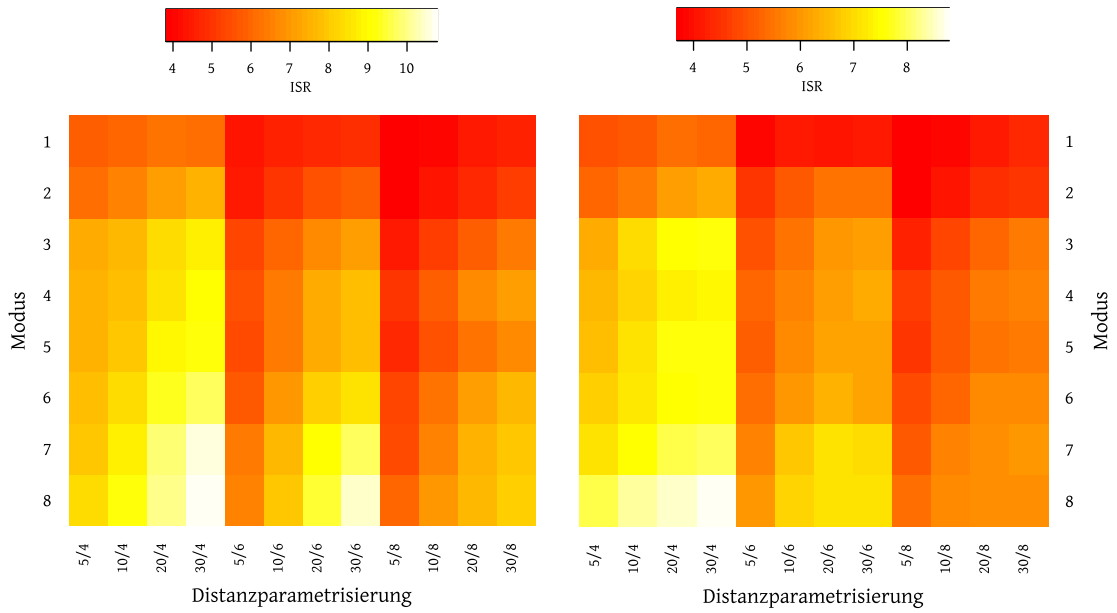
Zur Bestimmung der optimalen Form molekularer Kontextinformation wurden die möglichen Kombinationen der vorgestellten Gruppierungen für Nukleinsäuren und Proteine systematisch evaluiert. Der Vollständigkeit halber wurde dabei auch der Effekt überprüft, der beim Wegfall der atomaren Komponente aus der Beschreibung der Atomtypen eintritt. Insgesamt ergaben sich bei dieser Evaluation acht Modi für die Beschreibung der Atomtypen. Deren Überblick ist in Tabelle 5.2 gegeben.

Tab. 5.2: Übersicht über die verschiedenen Modi der kontextuellen Information. Zu jedem Modus ist neben seiner Identifizierungsnummer angegeben, ob die ursprünglichen Atomtypen in die Beschreibung einfließen und welche Art der Kontextinformation auf Seiten der Proteine und Nukleinsäuren verwendet wird.

Modus	Atomtypen	Protein	Nukleinsäure
1	nein	elektrostatisch	ja
2	nein	vollständig	ja
3	ja	nein	nein
4	ja	nein	ja
5	ja	elektrostatisch	nein
6	ja	elektrostatisch	ja
7	ja	vollständig	nein
8	ja	vollständig	ja

Distanzparametrisierung Die Autoren der SPA-PN-Potentiale berücksichtigten in ihrem Verfahren Interaktionen in einem Distanzbereich von $2{,}2$ Å bis $8{,}2$ Å mit einer Auflösung von 20 Zwischenschritten [342]. Während die untere Grenze durch das Auftreten erster Interaktionen aus dem verwendeten Datensatz hervorging, wurde weder die Wahl der oberen Grenze noch die verwendete Auflösung weiter durch die Autoren begründet. In vorhergegangen Studien herrschte Uneinigkeit über die Festlegung eines oberen Grenzwertes für die Berücksichtigung von molekularen Interaktionen. So ist im Zusammenhang mit Protein-Nukleinsäure-Komplexen gezeigt worden, dass kleine *Cutoff*-Werte die positive Diskriminierungsfähigkeit von kurzen Kontakten optimal zur Geltung bringen. Es wurde beobachtet, dass bereits ab Überschreitung eines oberen Grenzwertes von 6 Å eine signifikante Abnahme der Modellgüte zu verzeichnen war [339; 366]. Auch in einer davon unabhängigen Untersuchung, die sich *Cutoff*-Werten bis zu 15 Å widmete, wurde eine tendenziell niedrige Empfehlung für den oberen Grenzwert der Interaktionsdistanz ausgestellt. Interaktionen einer Länge von 6 Å bis 7 Å machten in diesem Fall ungewollte Nachbarschaftseffekte unvermeidlich [318]. Da Nachbarschaftseffekte jedoch eine Form der Kontextinformation darstellen können, stellte diese Argumentation keinen Grund für einen Ausschluss höherer Grenzwerte von der angestrebten Untersuchung dar. Weitere Studien kamen zu einem gegensätzlichen Ergebnis und erreichten auch mit *Cutoff*-Werten im Bereich von 10 Å bis 20 Å positive Ergebnisse [340; 344]. Auch im Protein-Protein-Umfeld gibt es unterschiedliche Meinungen über die Wahl des oberen Grenzwertes für die Berücksichtigung von atomaren Interaktionen. Es finden sich sowohl Studien mit einer niedrigen Empfehlung im Bereich von $3{,}5$ Å bis $6{,}5$ Å als auch Hinweise auf die bessere Wirksamkeit größerer Grenzwerte im Bereich größer als 10 Å [367; 368]. Die Festsetzung dieser Grenzwerte wurde aufgrund der unklaren Informationslage neu überprüft.

Zum Finden der optimalen Distanzparametrisierung wurde sowohl der obere Grenzwert für Interaktionen als auch die Auflösung im Distanzbereich überprüft. Die Modifikation der Werte fand dabei ausgehend von der Parametrisierung des Originalverfahrens in beide Richtungen statt. Bei einem oberen Distanz-*Cutoff* von unter $4{,}0$ Å relativ zur unteren Begrenzung war die Datenlage nicht ausreichend, um das Optimierungsverfahren durchzuführen. Für den *Cutoff* wurden folglich die Werte 4 Å, 6 Å und 8 Å verwendet. Bei der Distanzauflösung wurden neben dem Vorgabewert 20 noch die Alternativen Werte 5, 10 und 30 überprüft.



Trainingsdatensatz **(a)** **(b)** Testdatensatz

Abb. 5.7: Vergleich der Parametrisierung bezogen auf den erreichten ISR-Wert. Der ISR ist durch die Farbskala von niedrig (rot) bis hoch (weiß) auf die Parametrisierung aufgetragen. Neben den Modi (außen) wurden die beiden Distanzparameter (unten) in der Form Auflösung/Distanzbereich zusammengefasst.

Auswahl der optimalen Parametrisierung Für jede der aus den obigen Variationsbereichen mögliche Parameterkombination aus Distanzparametrisierung und Kontextinformation wurde der gesamte Generierungsprozess der SPA-PN Potentiale durchlaufen, um einen weitreichenden Vergleich zu ermöglichen. Als Abbruchkriterium wurde dabei das Erreichen des maximalen ISR für den Trainingsdatensatz definiert. Neben dem ISR wurde der Mittelwert der strukturellen Abweichung der jeweils am besten bewerteten Strukturen bestimmt. Wie bereits festgestellt, waren die charakteristischen Verläufe dieser beiden Kennwerte im Trainingsdatensatz stärker ausgeprägt als im Testdatensatz. Sie zeigten jedoch in beiden Datensätzen ähnlich positiven Tendenzen bezüglich des Optimierungserfolgs.

Aus dem Vergleich ging deutlich hervor, dass die Nutzung der Atomtypeninformation im Modus 3 gegenüber einer bloßen Betrachtung von Residuentypen in den Modi 1 und 2 einen Gewinn brachte. Die besten Ergebnisse wurden jedoch durch die Kombinationen dieser beiden Beschreibungstypen in Modi 4 bis 8 erreicht. Bezogen auf Grad der erreichbaren Spezifität der modifizierten SPA-PN-Potentiale korrelierte diese Verbesserung positiv mit der Menge der eingebrachten Information, wie aus dem Vergleich von Tabelle 5.2 mit Abbildung 5.7 hervorgeht. Entsprechend erreichten die Modi mit vollständiger Gruppierung der Aminosäuren sowohl bei Trainings- als auch bei Testdaten höhere Spezifitäten als diejenigen mit der elektrostatischen Gruppierung. Auch durch die Hinzunahme der Nukleotidtypen konnte ein Gewinn erzielt werden. Dieser war zwar im Trainingsdatensatz besonders bei der Anwendung der vollständigen Gruppierung kaum ausgeprägt, sein Wert zeigte sich jedoch bei der Betrachtung des gesonderten Testdatensatzes. Da die beschriebene Tendenz ebenso in der Entwicklung der gemittelten besten RMSD-Werte in Abbildung 5.8 erkennbar ist, wurde Modus 8 für die weitere Betrachtung präferiert. In Bezug auf die Distanzparametrisierung gehen zwei Trends aus dem Vergleich in Abbildung 5.7b hervor. Während das Einbeziehen weiter entfernter atomarer Kontakte zu einer deutlichen Verringerung der erzielbaren intrinsischen Spezifität führte, wurde mit einer höhe-

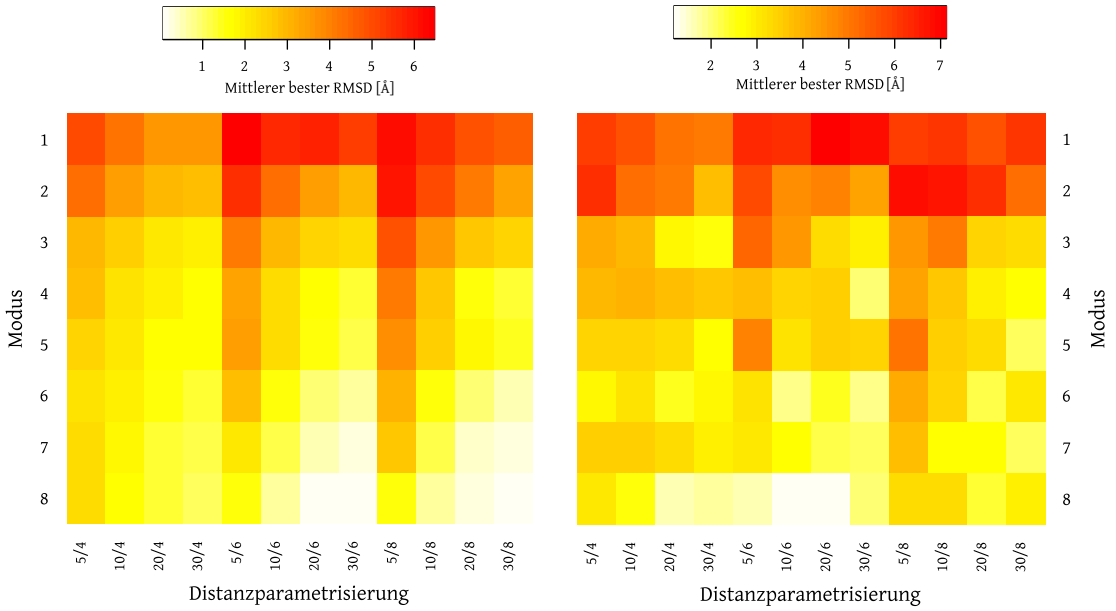


Abb. 5.8: Vergleich der Parametrisierung bezogen auf den gemittelten RMSD der am besten bewerteten Strukturen. Der RMSD ist durch die Farbskala von hoch (rot) bis niedrig (weiß) auf die Parametrisierung aufgetragen. Neben den Modi (außen) wurden die beiden Distanzparameter (unten) in der Form Auflösung/Distanzbereich zusammengefasst.

ren Auflösung über den betrachteten Distanzbereich in allen Fällen auch eine Erhöhung der intrinsischen Spezifität erreicht. Betrachtet man die Entwicklung der gemittelten besten RMSD des Trainingsdatensatzes in Abbildung 5.8, so ist zwar der Trend der Auflösung erkennbar, der Einfluss des Distanz-*Cutoffs* scheint aber entgegengesetzt. Eine Gegenprüfung mit den Ergebnissen des Testdatensatzes konnte dieses umgekehrte Verhalten jedoch nicht bestätigen, sodass an dieser Stelle von einer Überanpassung des Optimierungsschrittes auszugehen war. Zwar war im Testdatensatz die Distinktivität der RMSDs über die Distanzparametrisierung eher schwach ausgeprägt, es war jedoch erkennbar, dass die Vergrößerung des Distanzbereichs keinen Vorteil hatte. In Übereinstimmung mit den ISR-Bewertungen zeigte sich hier eine Präferenz für die Parametrisierungen mit einem Distanzbereich von 4 Å bis 6 Å und einer Auflösung über diesen Bereich von 20 bis 30 Abschnitten.

Im Vergleich hat sich gezeigt, dass eine hohe Auflösung in einem kleinen Distanzbereich unter Einbezug maximaler atomarer Informationen sowohl bei der ISR-Bewertung der Originalstrukturen als auch in Bezug auf die strukturelle Abweichung der am besten bewerteten *Decoy*-Kandidaten die optimalen Ergebnisse lieferte. Hierfür können zwei Gründe festgestellt werden. Dies ist zum Einen die grundsätzlich höhere Bedeutung der kurzstreckigen Interaktionen für die Ausbildung der molekularen Schnittstellen, da die Stärke einer Interaktion in der Regel umgekehrt mit deren Länge korreliert ist. Zum Anderen ist die Wahrscheinlichkeit für das Auftreten von Atompaaren ohne tatsächliche Wechselwirkungen in größeren Distanzbereichen wesentlich höher, wodurch Störgrößen in die Grunddatenmenge Einzug finden. Bei der Ableitung der Paarpotentiale führen diese Störgrößen dann zu einer Beeinträchtigung der Aussagekraft. Nach aktuellem Stand können sie nur durch die Verringerung der maximal betrachteten Distanz verhindert werden. Beim Festlegen der optimalen Auflösung ist neben den reinen Ergebnissen der Optimierungsläufe auch die Qualität der Eingangsstrukturen als Referenzmaß für die Bewertung der Relevanz zu beachten. Bei Strukturen, welche mittels XRD im mittleren bis hohen

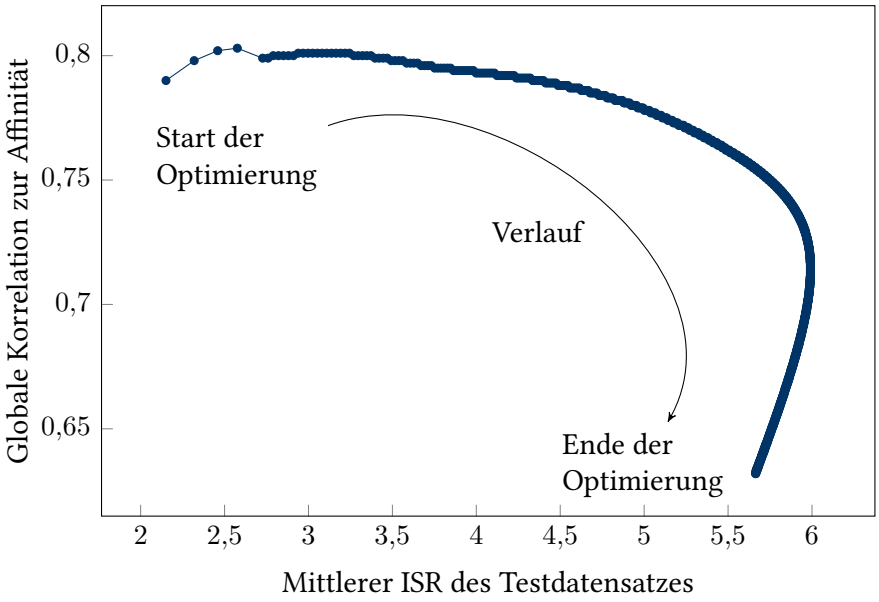


Abb. 5.9: Gegenläufigkeit der Spezifitätsbewertung (ISR) zur Korrelation zwischen Bewertung und freier Bindungsenergie während 1000 Iterationen der Optimierung der Paarpotentiale am Beispiel der unmodifizierten Parametrisierung.

Auflösungsbereich bestimmt wurden, ist typischerweise mit Abweichungen in den atomaren Koordinaten von bis zu 0,2 Å bis 0,3 Å zu rechnen [369; 370]. Eine Auflösung der distanzbasierten Paarpotentiale über diesen Fehlerbereich hinweg ist daher nicht zweckmäßig. Auch wenn dadurch nur der zweithöchste ISR erreicht wurde, ließ die Zweckmäßigkeitsgrenze nur die Auswahl der Distanzparametrisierung 4,0 Å / 20 im Modus 8 zu. Die 412 relevanten Atomtyppaare in 20 Abstandsbereichen lieferten für das modifizierte Paarpotential SPA-PN 8240 Einzelwerte.

Evaluation des Ergebnisses Die Autoren von SPA-PN haben die Wirksamkeit der Potentiale in der Vorhersage von Bindungsaffinitäten an einem Datensatz gezeigt, der bereits mehrfach [315; 318] verwendet wurde, tatsächlich jedoch eine Modifikation eines bereits eher definierten Datensatzes [337] von Protein-DNA-Strukturen darstellte. Der erreichte Korrelationskoeffizient von $>0{,}8$ zwischen freier Bindungsenergie und SPA-PN-Bewertung [342] konnte durch einen Teil der Parameterkombinationen nachvollzogen werden. Auffällig war dabei, dass die Entwicklung dieser Korrelationskoeffizienten gegenläufig zur Entwicklung der Spezifität ISR verlief, wie in Abbildung 5.9 dargestellt ist. Dies bedeutet, dass die nach ISR gesteuerte Optimierung bereits nach wenigen Schritten eine Abnahme der Vorhersagekraft für die Affinitäten bewirkte. Da sich dieses gegenläufige Verhalten der beiden Bewertungsgrößen bei allen relevanten Parameterkombinationen abzeichnete, wurden zur Verifikation weitere Datensätze herangezogen. Dies war zum einen der Ursprungsdatensatz mit 45 Komplexstrukturen, von denen durch fehlerhafte Bezeichner jedoch nur 42 auffindbar waren [337] und zum anderen ein Protein-RNA-Datensatz mit 73 Strukturen, von denen 72 genutzt werden konnten [371]. Zusätzlich zu diesen erprobten Datensätzen wurde aus Protein-DNA/RNA-Strukturen der Datenbank NPIDB [360] ein gemischter Datensatz mit Affinitätsannotation abgeleitet. Über experimentell bestimmte Dissoziationskonstanten k_D wurde dabei jeweils auf die freie Bindungsenergie ΔG geschlossen, um die Korrelation zu berechnen. Sowohl bei den beiden manuell selektierten Datensätzen als auch bei verschiedenen Teilmengen der automatisiert aggregierten Komplexstrukturen konnte durch keine der Parameterkombinationen eine Korrelation $>0{,}5$ erreicht werden. Die hervorra-

gende Korrelation aus der Originalpublikation scheint unter dem Licht dieser Gegenkontrolle das Ergebnis einer geschickten Auswahl des Testdatensatzes zu sein, nicht jedoch ein Indikator der Eignung dieser Paarpotentiale für die globale Abschätzung der Bindungsaffinität. Dies wird von einer Studie unterlegt, die an einem anderen Paarpotential eine ähnlich schlechte Eignung zur Affinitätsabschätzung feststellte [371]. Sowohl die Trainingsdaten als auch der gesamte Optimierungsprozess waren darauf ausgelegt, native von *Decoy*-Strukturen zu unterscheiden und damit eine relative Wertung der Strukturvarianten eines Komplexes vorzunehmen. Es war daher wenig überraschend, dass mit diesem Verfahren über viele unterschiedliche Strukturen hinweg keine globale Korrelation mit der Affinität erreichbar war.

5.2.3 Die ITScore-PR Potentiale

Wie auch bei den SPA-PN findet sich im Kern der Erzeugung der Potentiale ITScore-PR ein iterativer Optimierungsalgorithmus zur Verfeinerung der Ergebnisse. Im Gegensatz zu den SPA-PN fließt hier jedoch nicht die Spezifität in die Berechnung ein, sondern eine spezielle Bewertung mit Mitteln der statistischen Mechanik. Dies bildet nicht nur das Wesen der atomaren Interaktion realitätsgetreuer ab, es umgeht auch das Problem der Definition eines validen Referenzzustandes, welches in wissensbasierten Potentialen gelöst werden muss [344]. Die ITScore-PR sind daher als eine sinnvolle Ergänzung der bisherigen Verfahrenslandschaft zu betrachten, die in der Entwicklung hin zur molekularmechanischen Bewertung einen Zwischenschritt darstellt. Neben der Herleitung der Potentiale wurden die Ergebnisse in Form der konkreten Bewertungsfunktion und einer Referenzimplementierung vom Autor zur Verfügung gestellt [372].

Datengrundlage Zur Generierung eines Trainingsdatenbestandes wurden die Protein-RNA-Struktureinträge der PDB [305] von den Autoren entsprechend strenger Qualitätskriterien selektiert. Neben der Auflösung der Strukturaufklärung umfassten diese auch die Größe der Strukturen und Teilstrukturen. Einige Kriterien machten eine manuelle Kontrolle der Strukturen notwendig. Nach dem Bezug der ausgewählten Strukturen wurden mit Hilfe der Softwaresuite Rosetta [362] für jeden nativen Komplex 2000 *Decoy*-Strukturen erzeugt, von denen die 1000 besten Strukturen nach Bewertung durch ZDOCK [373] weiter verwendet wurden [344]. Die für die Ableitung der Potentiale relevanten Atomtypen bezogen die Autoren aus zwei Vorstudien [343; 374]. Atomare Interaktionen wurden bis zu einem oberen Schwellwert von $10\,\text{Å}$ in Schritten zu je $\Delta r = 0{,}2\,\text{Å}$ in die Berechnung einbezogen. Da durch die fehlende untere Begrenzung unbesetzte Schalen auftraten, wurde ein Maximalwert für die Bewertung festgelegt. Jede der 50 resultierenden Distanzschalen wird in der weiteren Vorstellung mit ihrem inneren Radius r angesprochen und erstreckt sich entsprechend der Schalenbreite bis $r + \Delta r$ [344].

Ableitung der initialen Potentiale Als Grundlage für die folgende Optimierung dienen die sogenannten initialen Potentiale $u_{ij}^{(0)}$. Nach der Berechnung aller Paarverteilungsfunktionen $g_{ij}^{k*}(r)$ der K nativen Komplexe werden aus diesen zunächst entsprechend Formel 5.9 die gemittelte Paarverteilungsfunktion $g_{ij}^{\text{obs}}(r)$ bestimmt. Die initialen Potentiale $u_{ij}^{(0)}(r)$ werden schließlich entsprechend Formel 5.11 abhängig vom Auftreten einer Wasserstoffbrückenbindung im betrachteten Paar unterschiedlich berechnet. Tritt eine Wasserstoffbrücke zwischen den zugehörigen Atomen auf, so kommt das einfach gemittelte Potential w_{ij} aus Formel 5.10

direkt zur Anwendung. Andernfalls setzt sich der konkrete Wert des Paarpotentials aus Beiträgen des gemittelten Potentials w_{ij} und des (6-12)-Lennard-Jones-Potentials v_{ij} zusammen, wie in Formel 5.11 näher spezifiziert wird [344].

$$g_{ij}^{\text{obs}}(r) = \frac{1}{K} \sum_{k=1}^{K} g_{ij}^{k*}(r) \tag{5.9}$$

$$w_{ij} = -k_B T \cdot \ln g_{ij}^{\text{obs}}(r) \tag{5.10}$$

$$u_{ij}^{(0)}(r) = \begin{cases} w_{ij}(r) & \forall \text{ Paare mit Wasserstoffbrücke} \\ \frac{v_{ij}(r) \cdot \mathrm{e}^{-v_{ij}(r)} + w_{ij}(r) \cdot \mathrm{e}^{-w_{ij}(r)}}{\mathrm{e}^{-v_{ij}(r)} + \mathrm{e}^{-w_{ij}(r)}} & \forall \text{ Paare ohne Wasserstoffbrücke} \end{cases} \tag{5.11}$$

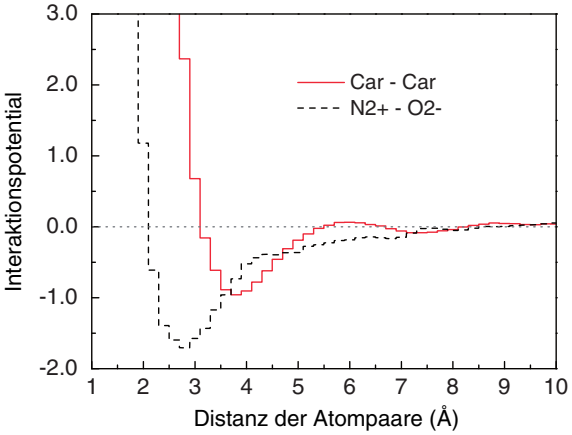
Ableitung der erwarteten Verteilungsfunktion Bei der Ableitung der erwarteten Verteilungsfunktion $g_{ij}^{(n)}(r)$ wird die Annahme zugrunde gelegt, dass die Menge der generierten *Decoy*-Strukturen $l \in L$ eine Boltzmann-Verteilung um die native Struktur k bildet. Die native Struktur ist dabei Teil des Ensembles und wird daher formal als 0. *Decoy*-Struktur behandelt. Die Paarverteilungsfunktionen $g_{ij}^{kl}(r)$ aller *Decoy*-Strukturen werden schließlich mit den Boltzmann-Wahrscheinlichkeiten P_k^l gewichtet und gemittelt, um die erwartete Verteilungsfunktion so zu erhalten, wie in Formel 5.12 angegeben. Die Boltzmann-Wahrscheinlichkeiten werden dabei aus den aktuellen Potentialen $u_{ij}^{(n)}(r)$ abgeleitet. Da diese im ersten Schritt, also bei Iteration 0, noch nicht definiert sind, treten an ihre Stelle die initialen Potentiale $u_{ij}^{(0)}(r)$ [344].

$$g_{ij}^{(n)}(r) = \frac{1}{K} \sum_{k=1}^{K} \sum_{l=0}^{L} P_k^l g_{ij}^{kl}(r) \tag{5.12}$$

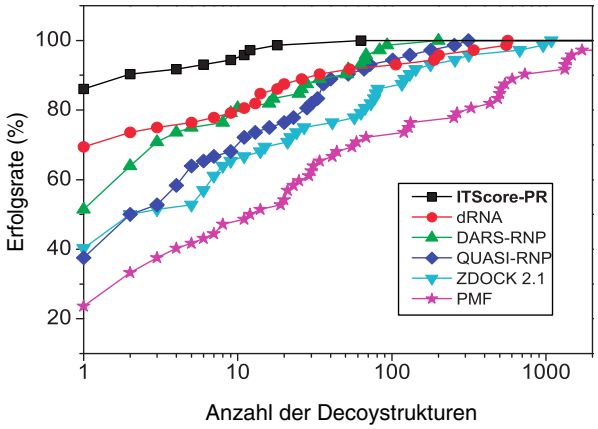
Iterative Optimierung Die finalen Atompaarpotentiale werden ebenfalls durch einen iterativen Prozess schrittweise optimiert. Dazu wird in jeder Iteration n die Differenz zwischen der erwarteten Verteilungsfunktion der aktuellen Iteration und der beobachteten Verteilungsfunktion bestimmt und entsprechend Formel 5.13 durch einen konstanten Faktor in ihrer Größenordnung reguliert. Die erwarteten Potentiale der Folgeiteration ergeben sich anschließend wie in Formel 5.14 angegeben durch Korrektur der aktuellen Potentiale mit der ermittelten Differenz. Da die erwarteten Potentiale über die Boltzmann-Wahrscheinlichkeiten P_k^l in die Berechnung der erwarteten Verteilungsfunktion $g_{ij}^{(n)}$ einfließen, entsteht die im Optimierungssystem notwendige Rückkopplung [344]. Als erstes Ergebnis zeigt Abbildung 5.10a, dass die erhaltenen Potentiale die grundlegenden Charakteristika der bekannten physikalischen Potentiale aufweisen. Im Vergleich zu einigen anderen wissensbasierten Potentialen erzielen die ITScore-PR dabei gute Ergebnisse in der Bewertung von *Decoy*-Strukturen, wie die aus der Originalpublikation der Autoren stammende Abbildung 5.10b belegt.

$$\Delta u_{ij}^{(n)}(r) = \frac{1}{2} k_B T \left(g_{ij}^{(n)}(r) - g_{ij}^{\text{obs}}(r) \right) \tag{5.13}$$

$$u_{ij}^{(n+1)}(r) = u_{ij}^{(n)}(r) + \Delta u_{ij}^{(n)}(r) \tag{5.14}$$



(a) Verlauf zweier herausgegriffener Paarpotentiale der ITScore-PR, aufgetragen über die atomare Distanz.



(b) Erfolgsrate von ITScore-PR bei der Bewertung von *Decoy*-Strukturen im Vergleich zu anderen wissensbasierten Potentialen.

Abb. 5.10: Die Abbildungen aus der Originalpublikation von ITScore-PR zeigen exemplarisch den Potentialverlauf und die Überlegenheit der ITScore-PR Potentiale im Vergleich [344].

Bezug und Anwendung Die finalen ITScore-PR-Potentiale wurden auf der Internetpräsenz der Autoren nach schriftlicher Bestätigung der zugehörigen Lizenzvereinbarung zum Download angeboten [372]. Der Download enthielt neben den Rohdaten der Bewertungsfunktion eine kleine Anwendung, die es ermöglicht, Protein-RNA-Komplexe mittels ITScore-PR zu bewerten. Dabei ist zu beachten, dass die beiden Komplexpartner im PDB-Format zwar in separaten Dateien jedoch in einem gemeinsamen Koordinatensystem vorliegen müssen. Der Anwendung lag der Quelltext in der Programmiersprache C bei, sodass eine Portierung nach Java möglich war. Da DNA und RNA im molekularen Aufbau eine sehr hohe Ähnlichkeit aufweisen, bestand die Vermutung, dass die auf Komplexstrukturen von Protein und RNA angelernte Bewertungsfunktion ebenfalls auf Protein-DNA-Komplexe anwendbar war. Beim entsprechenden Aufruf der Software zeigte sich jedoch deren Beschränkung auf das Einlesen von Protein- und RNA-Ketten. In der eigenen Implementierung wurde daher die Eingaberoutine so angepasst, dass auch DNA-Stränge verarbeitet werden konnten. Dazu wurden zunächst die DNA-Residuenbezeichner eingeführt und deren atomare Definition sichergestellt. Das bei DNA fehlende Sauerstoffatom an der $2'$-Position des Pentoserings erforderte keinen Handlungsbedarf. Für die hinzugekommene Methylgruppe an Position 5 des Pyrimidinringes beim Übergang von Uracil zu Thymin wurde anhand der vorhandenen Atomtypen eine entsprechende Definition des Kohlenstoffatoms eingefügt. Der Versuch zeigte, dass die so modifizierte Routine in der Lage war, auch Protein-DNA-Komplexe zu bewerten. Von der Verwendung der optionalen Möglichkeit zur geringfügigen Optimierung der Geometrie über ein nicht-deterministisches Simplex-Verfahren wurde im weiteren abgesehen, um die Vergleichbarkeit der Ergebnisse zu gewährleisten.

5.2.4 Molekularmechanische Bewertung

Nach Annahme der entsprechenden Lizenzvereinbarungen konnte sowohl die HADDOCK-Hauptapplikation [375] als auch das Softwarepaket AmberTools16 [376] von den Internetpräsenzen der Hersteller bezogen werden.

AMBER Die molekularmechanische Bewertung über das Kraftfeld AMBER basiert sowohl auf experimentellen als auch auf simulierten Daten. Grundlegend fließen in die Bewertung, wie sie in Formel 5.15 umrissen wird, vier Terme ein. Bindungslängen und -winkel werden dabei wie ein harmonischer Oszillator behandelt, ein durch seine lineare Rückstellgröße ausgezeichnetes, schwingungsfähiges System. Die für die Parametrisierung dieser Terme notwendigen Analoga zu Rückstellgröße (r_{eq}, θ_{eq}) und Kraftkonstante (K_r, K_θ) stammen ähnlich wie die Parametrisierung des Torsionswinkelterms aus Strukturdaten und Schwingungsspektren kleiner Moleküle. Der vierte Term charakterisiert nicht-kovalente Interaktionen über van-der-Waals- und elektrostatische Wechselwirkungen spezifischer Atompaare. Grundlage für die Parametrisierung des Lennard-Jones- (A_{ij}, B_{ij}) und des Coulomb-Potentials (q_i, q_j) bilden dabei Monte-Carlo-Simulationen von kurzen Kohlenwasserstoffen sowie quantenmechanische Berechnungen [346]. Über die ursprüngliche Variante des Kraftfeldes ff94 hinweg gab es eine Reihe von Korrektur- und Optimierungsversuchen, welche in weiteren Versionen des Kraftfeldes resultierten. Die aktuellste ist `ff14SB` [377]. Ein Vergleich an den Versionen `ff94`, `ff99`, `ff10`, und `ff14SB` zeigte zwar deutliche Veränderungen der konkreten Energiebewertungen, die relative Charakteristik der Beschreibungen zueinander blieb jedoch zwischen den Versionen des Kraftfelds nahezu konstant.

Die Energiebewertung mit dem aktuellen Kraftfeld AMBER `ff14SB` erforderte eine Vorverarbeitung der Molekülkomplexe mit anschließender Überführung in ein kompatibles Format. Nach Auftrennung der Komplexstruktur in die beiden Bindungspartner konnten schließlich unter Zuhilfenahme der Software *Multiscale Modeling Tools for Structural Biology* (MMTSB) [378] sowohl die energetische Bewertung der ungebundenen Komplexpartner als auch die ihrer gebundenen Form durch die Suite AmberTools16 [376] bestimmt werden. Die energetische Bewertung der vorliegenden Bindung ergab sich schließlich aus der Differenz dieser beiden Werte.

$$\begin{aligned} E_{tot} = & \sum_{\text{Bindungen}} K_r(r - r_{\text{eq}})^2 + \sum_{\text{Winkel}} K_\theta(\theta - \theta_{\text{eq}})^2 \\ & + \sum_{\text{Diederwinkel}} \frac{V_n}{2}\left(1 + \cos(n\phi - \gamma)\right) + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right] \end{aligned} \tag{5.15}$$

HADDOCK Die Bewertungsfunktion der Dockingsuite HADDOCK setzt sich aus einer Reihe von Termen zusammen, die vom Nutzer für jede Dockingphase separat gewichtet werden können. Neben der Charakterisierung nicht-kovalenter Interaktionen durch die entsprechenden Terme des OPLS-Kraftfeldes gehören auch ein spezieller Desolvationsterm und ein Maß für die Fläche der entstandenen Schnittstelle zur Bewertung. Zusätzlich existieren für die verschiedenen Arten von Informationen, die in den informationsgetriebenen Dockingprozess eingebracht werden können, weitere Terme, die die Erfüllung der dadurch spezifizierten Bedingungen erfassen [352]. Unter Auslassung von nicht-gewichteten und nicht-relevanten Termen ergeben sich in HADDOCK die Standardgewichtungen entsprechend Tabelle 5.3. Besonders auffällig ist dabei, dass der Einfluss der Elektrostatik in den beiden unsolvatisierten Phasen beim Übergang in die solvatisierte Phase stark zugunsten der van-der-Waals-Wechselwirkungen verringert wird. So wird beim Finden der initialen Konstellationen neben dem Desolvationsterm quasi nur die Elektrostatik berücksichtigt, während im Zuge der finalen Verfeinerung die van-der-Waals-Wechselwirkungen und Abweichungen von den vorgegebenen Distanzkriterien wesentlich stär-

Tab. 5.3: Gewichtungsschemata der HADDOCK-Bewertungsfunktion mit den Energietermen für van-der-Waals-Interaktionen (`vdw`), Elektrostatik (`elec`), AIR (`dist`), Fläche der Schnittstelle (`bsa`) und dem Desolvationsterm (`desolv`).

Dockingphase	`vdw`	`elec`	`dist`	`bsa`	`desolv`
Unsolvatisierte Phase 1	0,01	1,00	0,01	-0,01	1,00
Unsolvatisierte Phase 2	1,00	1,00	0,10	-0,01	1,00
Solvatisiert Phase	1,00	0,20	0,10	0,00	1,00

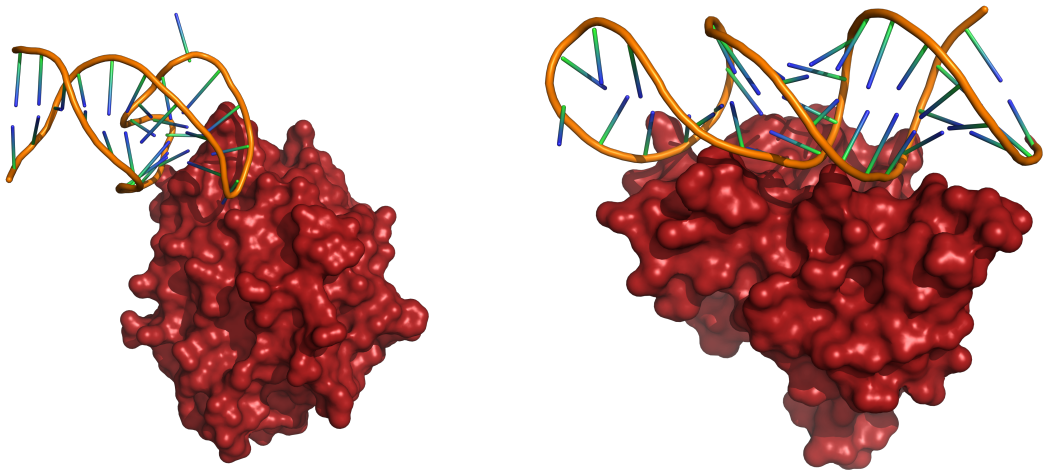
ker eingehen. Die Fläche der ausgebildeten Schnittstelle hingegen findet in der initialen Phase kaum Beachtung und wird aus der Bewertung während der Verfeinerungsphase schließlich ganz ausgelassen.

5.3 Konzeption des Vergleichs der Bewertungsmodelle

Aufgrund der bereits erfolgten Implementierung lag nahe, die Bewertung nach ISR als Kriterium für den Vergleich der vorgestellten Bewertungsmodelle einzusetzen. Als Optimierungskriterium der SPA-PN hätte dies jedoch die Neutralität der Untersuchung beeinträchtigt. Tatsächlich zeigte ein solcher Vergleich die unübertreffliche Eignung der SPA-PN. Allein auf den ITScore-PR angewendet zeigte der ISR als Vergleichskriterium jedoch mit geringem Aufwand, dass sowohl für DNA als auch für RNA als Bindungspartner ähnliche Wertungen erreicht werden können. In dem gemischten Datensatz, der zur Evaluation der SPA-PN verwendet wurde, erreichten die ITScore-PR für die intrinsische Spezifität Werte von $3{,}2$ bei RNA beziehungsweise $2{,}9$ bei DNA. Auch wenn die Abweichung beträchtlich erschien, befand sie sich in Betracht der geringen Größe des Datensatzes sowie der strukturellen und damit physikochemischen Diversität der Strukturen im erwarteten Bereich. Dies gab einen ersten Hinweis darauf, dass die Bewertung nach ITScore-PR wirklich mit beiden Nukleinsäuretypen kompatibel war. Für einen verlässlichen Vergleich zwischen den vier Bewertungsmodellen wurde jedoch eine unabhängige und neutrale Methodik erforderlich. Zu diesem Zweck wurden zwei Referenzkomplexe mit deutlich unterschiedlichen Eigenschaften gewählt, die nicht Teil der Trainingsmengen der vier Bewertungsmodelle waren. Unter Zuhilfenahme einer semi-flexiblen Dockingsimulation wurden für diese zwei Komplexe schließlich *Decoy*-Strukturen erzeugt und durch die Modelle bewertet. Die Wirksamkeit der Bewertungsmodelle zeigte sich dabei durch ihre Fähigkeit, native und nahenative Komplexe von nicht-nativen zu unterscheiden.

5.3.1 Auswahl und Vorstellung der Referenzkomplexe

Für den Vergleich wurde sowohl für DNA als auch für RNA als Bindungspartner jeweils eine experimentell aufgeklärte Protein-Aptamer-Komplexstruktur aus der Datenbank PDB [305] als Referenz gewählt. Dabei wurde durch manuelle Überprüfung der jeweiligen Trainings- und Testdatensätze sichergestellt, dass die Referenzstrukturen nicht an den Erzeugung und internen Verifikation der beiden wissenschaftlichen Potentiale beteiligt waren. Es wurde ferner auf die Unterschiedlichkeit der von ihnen ausgebildeten Bindeflächen geachtet. Die Referenzkomplexe sind in Abbildung 5.11 dargestellt.



5CMX [379] **(a)** **(b)** 4PDB [380]

Abb. 5.11: Darstellung der Referenzkomplexe, die zur Verifikation genutzt wurden. Das Zielprotein wird mit seiner lösungsmittelzugänglichen Oberfläche in roter Färbung dargestellt. Für das Aptamer wird die Cartoon-Darstellung verwendet. Die beiden Komplexe unterscheiden sich sowohl in der relativen Lage der beteiligten Partner als auch im Flächeninhalt der molekularen Schnittstelle.

Die Struktur `5CMX` Bei der ersten Komplexstruktur handelt es sich um ein DNA-Aptamer, welches an humanes sowie bovines Thrombin bindet. Durch die Komplexbildung mit dem Gerinnungsfaktor Thrombin wird dessen Wirkung im Blutgerinnungsprozess inhibiert, was einen Einsatz des Aptamers als Antikoagulans ermöglicht. Basierend auf den Ergebnissen vorheriger SELEX-Läufe gegen das Zielmolekül Thrombin nutzten die Autoren ein $15\,\mathrm{nt}$ langes Sequenztemplate zur Synthetisierung einer Anfangsbibliothek. Auf diese Bibliothek wendeten sie schließlich das Screeningverfahren *Evolution-mimicking algorithm* (EMA) an, welches im Gegensatz zu SELEX nicht die Affinität des Aptamers, sondern dessen Enzyminhibition als Zielgröße nutzt. Die erfolgreichsten Kandidaten dieses Screeninglaufes wurden durch beidseitiges Flankieren auf eine Länge von $31\,\mathrm{nt}$ erweitert und erneut einem EMA-Screening unterzogen. Eines der wenigen positiven Resultate dieser Modifikation wurde unter dem Namen `RE31` weitergeführt [381]. Es konnte gezeigt werden, dass die inhibierende Wirkung des `RE31`-Aptamers speziesabhängig und bezogen auf humanes Thrombin anderen existierenden Anti-Thrombin-Aptameren deutlich überlegen war [382]. Die Komplexstruktur zwischen `RE31` und Thrombin wurde durch XRD mit einer Auflösung von $2{,}98\,\text{Å}$ aufgeklärt. Sie zeigt zum einen die bereits vermutete Duplex-Quadruplex-Faltung des Aptamers und zum anderen, dass diese Quadruplex-Formierung an die *Exosite I* des Thrombins bindet [383]. Die damit verbundene Blockierung der Bindungsstelle von Fibrinogen führt zur inhibierenden Wirkung des Aptamers [384]. Die finale Komplexstruktur wurde von den Autoren unter der PDB-ID `5CMX` zur Verfügung gestellt [379].

Die Struktur `4PDB` Bei der zweiten Komplexstruktur handelt es sich um ein RNA-Aptamer, welches an das ribosomale Protein S8 des *Bacillus Anthracis* bindet, dem Erreger von Milzbrand aus der Familie der *Bacillaceae*. Als eines der 21 Proteine der kleinen 30S-Untereinheit wirkt das ribosomale Protein S8 regulierend in der Translation von mRNA. Für das SELEX-Screening bereiteten die Autoren eine Startbibliothek so vor, dass die entstehenden Aptamere eine *Hairpin*-Struktur ausbildeten. Die verwendete Template-Sequenz ließ an zwei Stellen insgesamt 16

Tab. 5.4: Sequenzen des Anti-S8-Aptamers, sowohl in der vollständigen als auch in der verkürzten Variante in alignierter Form. Das Alignment selbst zeigt neben identischen (=) und nicht übereinstimmenden (*) Positionen, die entstehenden Lücken (-). Da im Strukturvergleich keine Lücken berücksichtigt werden können, gibt der geschlossene Konsensus den Teil der Sequenz an, der dem Strukturvergleich ohne Lücken zugänglich ist.

Name	Sequenz
Vollständiges Aptamer	GGGAUGCUCAGUGAUCCUUCGGGAUAUCAGGGCAUCCC
Gekürztes Aptamer	GGG.....CAGUGAUGCUUCGGCAUAUCAG.....CCC
Alignment	================*=======*================
Geschlossener Konsensus	........CAGUGAU.CUUCGG.AUAUCAG........

variable Nukleotide zur Randomisierung der Bibliothek zu. Nach zehn Runden wurde das entsprechende $42\,\mathrm{nt}$ lange Aptamer selektiert. Die Komplexstruktur aus Aptamer und bakteriellem S8-Protein wurde durch XRD mit einer Auflösung von $2{,}60\,\text{Å}$ aufgeklärt. In der Struktur ist zu erkennen, dass das Aptamer spezifisch an der 16S Ribosomale RNA (rRNA)-Bindestelle des Proteins fixiert ist. Die rRNA bindet an dieser Stelle über ein konserviertes Sequenzmotiv, das jedoch im betrachteten Aptamer nicht vorhanden ist. Obwohl Sequenzmotiv und Topologie des Aptamers nicht dem der rRNA-Bindestelle entsprechen, sind die auftretenden Interaktionen denen des nativen rRNA-S8-Komplexes durchaus ähnlich. Die finale Komplexstruktur wurde von den Autoren unter der PDB-ID 4PDB zur Verfügung gestellt [380].

Neben der strukturellen Aufklärung des Protein-Aptamer-Komplexes untersuchten die Autoren auch eine verkürzte Variante des Aptamers, siehe dazu Tabelle 5.4. Sie stellten dabei fest, dass die konformationelle Anpassung des Aptamers an seinen Bindepartner sehr umfangreich war [385], stellten jedoch nur die ungebundene Struktur des gekürzten Aptamers unter der PDB-ID 2LUN zur Verfügung [386]. Die Datenlage für die Verwendung des gekürzten Aptamers war schlecht, da eine Referenzstruktur in gebundener Form nicht zur Verfügung stand und aus einem Alignment zwischen den beiden Aptamervarianten mangels struktureller Ähnlichkeit nicht abgeleitet werden konnte. Es fehlte daher nicht nur die Bindungsinformation für die gezielte Erzeugung nahe-nativer *Decoy*-Strukturen, sondern auch die Basis für die Berechnung der strukturellen Abweichungen der Decoystrukturen. Als relevante Residuen verblieben in der vergleichenden Betrachtung ausschließlich diejenigen, die sich aus der Schnittmenge der an der Bindung beteiligten Reste aus Tabelle 5.5 und dem geschlossenen Konsensus aus Tabelle 5.4 ergaben. Aus der Untersuchung der verkürzten Aptamerstruktur ergab sich kein verwertbares Ergebnis bezüglich der Güte der Bewertungsfunktionen.

5.3.2 Generierung von Decoy-Strukturen

Zur Generierung von *Decoy*-Strukturen wurde eine semi-flexible Dockingsimulation der beiden Komplexpartner durchgeführt. Genutzt wurde dazu die Desktop-Version der Software HADDOCK in Version 2.2 [352; 387]. Die eigentlich für Protein-Protein-Docking entwickelte Software kann durch entsprechende Parametrisierung auch Dockingsimulationen durchführen, an denen Nukleinsäuren beteiligt sind. Die Einstellungen müssen zwar von Hand eingepflegt werden, werden jedoch bereits in der Standardinstallation unterstützt [375]. Um den Eingabespezifikationen von HADDOCK zu genügen, wurden die Referenzstrukturen neu indiziert und die Residuen entsprechend der 3-Letter-Namenskonvention benannt. Um sicherzustellen, dass sich

alle erzeugten *Decoy*-Strukturen tatsächlich von der jeweiligen Originalstruktur unterscheiden, wurde diese vor Beginn der Prozedur in ihre Einzelkomponenten aufgetrennt und durch eine manuelle Translation und Rotation der Nukleinsäure verändert.

Verwendung von Einschränkungen In der Vorbereitung einer informationsgetriebenen Dockingsimulation mit HADDOCK ist es möglich und gewollt, verschiedene Nebenbedingungen vorzugeben, die die Simulation zielführend einschränken. Während des Dockingprozesses werden Verstöße gegen die eingeführten Nebenbedingungen in den zugehörigen Termen der Bewertung bestraft, sodass sich Strukturvarianten durchsetzen, welche konform mit ihnen sind. Neben vielen weiteren experimentell abgeleiteten Nebenbedingungen sind im vorliegenden Fall hauptsächlich die sogenannten *Ambiguous Interaction Restraints* (AIR) relevant. Sie bilden über die Definition von aktiven und passiven Residuen sowie die Spezifikation einer einzuhaltenden Maximaldistanz der Kontakte bekannte oder vermutete Bindungsregionen des Zielkomplexes ab. Stehen keine derartigen Informationen zur Verfügung, so besteht die Möglichkeit, die AIR während der initialen Phase des Dockings jeweils zufällig aus den oberflächenzugänglichen Residuen der beiden Komplexpartner zu erzeugen. Mit hinreichend großer Zahl von Kandidatenstrukturen können damit nahezu alle relevanten Konstellationen der Bindungspartner abgedeckt werden. Die Dokumentation der Software HADDOCK empfiehlt die Aktivierung einer Reihe weiterer Nebenbedingungen, welche hauptsächlich über Aspekte der Oberfläche (*Surface Restraints*) und Massenmittelpunkte (*Center of Mass Restraints*) die Kompaktheit der erzielten Komplexstrukturen fördern sollen. Ein Testlauf mit diesen beiden Nebenbedingungen zeigte jedoch unter 5000 generierten Kandidatenkomplexen keine nahe-native Struktur, während ohne ihre Anwendung auch im nahe-nativen Bereich zahlreiche Komplexe generiert wurden. Für den vorliegenden Fall wurde daher aus mangelnder Eignung von ihrer Verwendung abgesehen, sodass ausschließlich zufällig generierte sowie aus den bekannten Strukturen abgeleitete AIR zum Einsatz kamen.

Dockingprotokoll Das angewendete Dockingprotokoll gliederte sich in drei aufeinanderfolgenden Phasen. In der ersten Phase wurden durch Translation und Rotation unter Beachtung der AIR zufällige Kombinationen der beiden Komplexpartner erstellt. Nach einer mehrstufigen Energieminimierung der Komplexe, in denen die jeweiligen Partner als starre Körper behandelt wurden, erfolgte anhand der HADDOCK-internen Bewertungsfunktion die Auswahl derjenigen Strukturen mit den besten intermolekularen Energiebewertungen. In der zweiten Phase wurde diese Teilmenge einer mehrstufigen Optimierung nach dem Prinzip der simulierten Abkühlung (*Simulated Annealing*) unterzogen, welche in den späteren Stufen leichte Konformationsänderungen der beiden Bindepartner zuließ. Im Anschluss erfolgte eine Energieminimierung nach dem Gradientenabstiegsverfahren. In der letzten Phase fand nach dem Prinzip der molekulardynamischen Simulation über einen sehr kurzen simulierten Zeitbereich eine Verfeinerung der Komplexstrukturen unter Einfluss expliziter Solvatisierung statt [352].

Um einen Datensatz zu erhalten, der sowohl nahe-native als auch deutlich abweichende Strukturen in ausreichender Menge enthält, wurden zwei unabhängige Durchläufe des Dockings durchgeführt und deren Ergebnisse vereint. Der erste Durchlauf erzeugte dabei mit ausschließlich zufällig erzeugten AIR quasi *ab initio Decoy*-Strukturen, die in einem großen Variationsbereich lagen. Der zweite Durchlauf nutze die bekannten Schnittstellen der nativen Strukturen als Orientierung. Dazu wurden aus den Beschreibungen der zugehörigen Publikationen und Strukturdateien die in Tabelle 5.5 aufgeführten aktiven Residuen bestimmt. Diese wurden aufge-

Tab. 5.5: An der molekularen Schnittstelle beteiligte Residuen der beiden Referenzkomplexe mit Distanzbereich der Interaktionen, wie sie aus der entsprechenden Publikation/Strukturdatei entnommen wurden [380; 383]

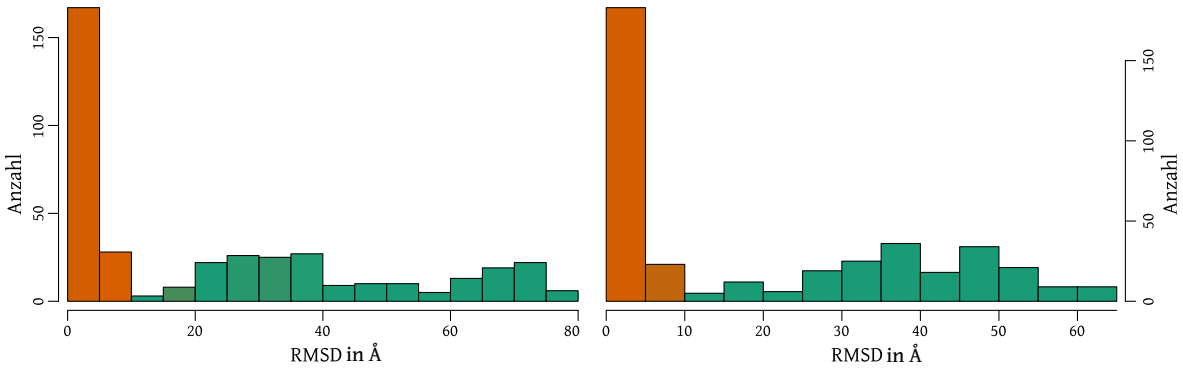
Komplex	Residuen der Schnittstelle	Distanzbereich
5CMX	T_{11}, T_{12}, G_{13}, T_{20}, T_{21}	2,28 Å bis 3,49 Å
	ARG_{75}, TYR_{76}, GLU_{77} ARG_{77A}, ASN_{78}, TYR_{117}	
4PDB	A_4, G_6, C_7, C_{16}, C_{17}, A_{24}, U_{25}, A_{26}, U_{27}	2,33 Å bis 3,30 Å
	LYS_{54}, GLN_{80}, ALA_{114}, SER_{130}, THR_{131} SER_{132}, LYS_{133}, THR_{146}, GLY_{147}, GLU_{149}	

arbeitet als AIR in den zweiten Dockinglauf eingebracht. Durch das Zulassen einer maximalen Distanz zwischen den Residuen der Schnittstelle von 3,5 Å bei der Struktur `5CMX` sowie 4,0 Å bei der Struktur `4PDB` wurde sichergestellt, dass genügend Flexibilität zur Ausbildung einer hohen Diversität unter den nahe-nativen *Decoy*-Strukturen vorhanden war. Beide Durchläufe erzeugten jeweils 1000 *Decoy*-Strukturen in der ersten Phase, von denen die 200 am besten bewerteten die Folgephasen durchliefen, sodass ein kombinierter Datensatz mit 400 diversen Strukturen entstand. Die HADDOCK-interne Bewertung der Strukturen konnte direkt aus den Metadaten der Simulation ausgelesen werden. Die Bewertungen nach AMBER, ITScore-PR und SPA-PN wurden schließlich unter Zuhilfenahme der bereits genannten Software berechnet und tabellarisch aufgearbeitet.

5.3.3 Quantifizierung der strukturellen Abweichung

Die mittlere strukturelle Abweichung zwischen der nativen Struktur und einem *Decoy*-Komplex wurde über den RMSD quantifiziert. Bei der Betrachtung dieses Kennwertes ließen sich zwei Bereiche, die nahe- und nicht-nativen Decoystrukturen unterscheiden.

Bestimmung des RMSD Die Bestimmung des atomaren Abstandes zweier Vergleichskomplexe über den RMSD erforderte durch die Bewegungen des Systems während der Dockingsimulation einen zusätzlichen Vorbereitungsschritt. Die Vergleichsstrukturen wurden dazu über ein semi-globales, strukturelles Alignment neu im Raum aneinander ausgerichtet. Da nur der Proteinbindepartner für diese Ausrichtung verantwortlich war, konnte die Nukleinsäure in der neuen Anordnung als alleiniger Träger der relevanten strukturellen Abweichung angesehen werden. Zwar wirkte sich die strukturelle Veränderung aus der flexiblen Dockingphase auch auf den alignierten Proteinbestandteil aus, der Einfluss war jedoch sehr klein und nicht systematisch, sodass er in der Betrachtung tatsächlich unbedeutend geblieben ist. Die Berechnung des atomaren Abstandes erfolgte anschließend über die Distanzen aller relevanten Atome der Nukleinsäure. Durch die Reduzierung der Betrachtung auf die Kohlenstoffatome wurde eine stabilere Beschreibung der strukturellen Gesamtabweichung erreicht. Die Kohlenstoffatome spiegeln gut die Gesamtkonformation der Nukleinsäure wieder ohne jedoch auf vernachlässigbare Fluktuationen übermäßig anzusprechen, da sie als strukturelle Basis sowohl im Rückgrat als auch in den Nukleobasen vorkommen.



5CMX **(a)** **(b)** 4PDB

Abb. 5.12: Verteilung der strukturellen Abweichungen von der nativen Komplexstruktur innerhalb der erzeugten *Decoy*-Strukturen mit (orange) und ohne (grün) Verwendung von Informationen über die native Schnittstelle. Eine entsprechende schwache Trübung der Farbe im Grenzbereich zwischen nahe- und nicht-nativem Sektor gibt Hinweis auf eine sehr geringfügige Überschneidung.

Definition der Distanzbereiche Die Verteilung der konkreten RMSD-Werte der *Decoy*-Strukturen spiegelte induziert durch die zweiteilige Erzeugung die erwarteten Distanzbereiche wider. Im nahe-nativen Bereich lagen die beiden Bindungspartner in einer Konformation vor, welche der nativen in Position und Ausrichtung sehr ähnlich war. Er erstreckte sich im RMSD-Bereich bis zu $10\,\text{Å}$, wobei die besten Komplexe Abweichungen von knapp unter $1\,\text{Å}$ erreichten. Ein Teil der Abweichung wurde durch die strukturelle Flexibilität innerhalb der Bindepartner hervorgerufen, welche in den späten Phasen der Dockingsimulation vom System eingeführt wurde. Im nicht-nativen Bereich befanden sich diejenigen *Decoy*-Strukturen, bei denen die Position oder Ausrichtung der molekularen Schnittstelle wesentlich von der nativen Vorgabe abwich. Beginnend mit Abweichungen über $10\,\text{Å}$ erstreckte sich der nicht-native Bereich für die *Decoy*-Strukturen des Referenzkomplexes `4PDB` bis zu $65\,\text{Å}$ und für die des Referenzkomplexes `5CMX` auf bis zu $80\,\text{Å}$. Die Auftragung der absoluten Häufigkeiten über definierte RMSD-Bereiche in Abbildung 5.12 zeigt eine deutliche Häufung im nahe-nativen Bereich. Sie ist das erwartete Ergebnis der vorherig beschriebenen Anordnung und umfasst etwa $50\,\%$ des *Decoy*-Ensembles. Wie in der detaillierteren Aufschlüsselung der Abbildungen 5.13 und 5.14 noch deutlicher erkennbar ist, wurde der nahe-native Bereich fast ausschließlich durch die Strukturen repräsentiert, die unter Anwendung der AIR entstanden. Ohne diese Zusatzinformation wurde im Gegenzug bis auf wenige Ausnahmen der nicht-native Bereich abgedeckt.

Aussagekraft der Quantifizierung Die manuelle Überprüfung einer Stichprobe der *Decoy*-Strukturen aus dem nahe-nativen Bereich ergab eine gute Eignung des RMSD zur quantitativen Abbildung des tatsächlich beobachteten, qualitativen Grades der strukturellen Verschiedenheit von der nativen Konformation. Die analoge Überprüfung bestätigte im nicht-nativen Bereich, dass der RMSD die große Vielfalt qualitativer Verschiedenheit hier nicht beschreiben konnte, da sich seine qualitative Bedeutung durch die zunehmende Unbestimmtheit des eindimensionalen Maßes bei größeren Abweichungen verliert. Diese Beeinträchtigung war jedoch im betrachteten Fall unerheblich, da zum Finden der nativen Konformation nicht die Art der Unähnlichkeit sondern der Grad der Ähnlichkeit von Interesse war. Dieser wurde im nahe-nativen Bereich hinreichend durch den RMSD beschrieben, sodass sich dieser als Bewertungskriterium für den folgenden Vergleich eignete.

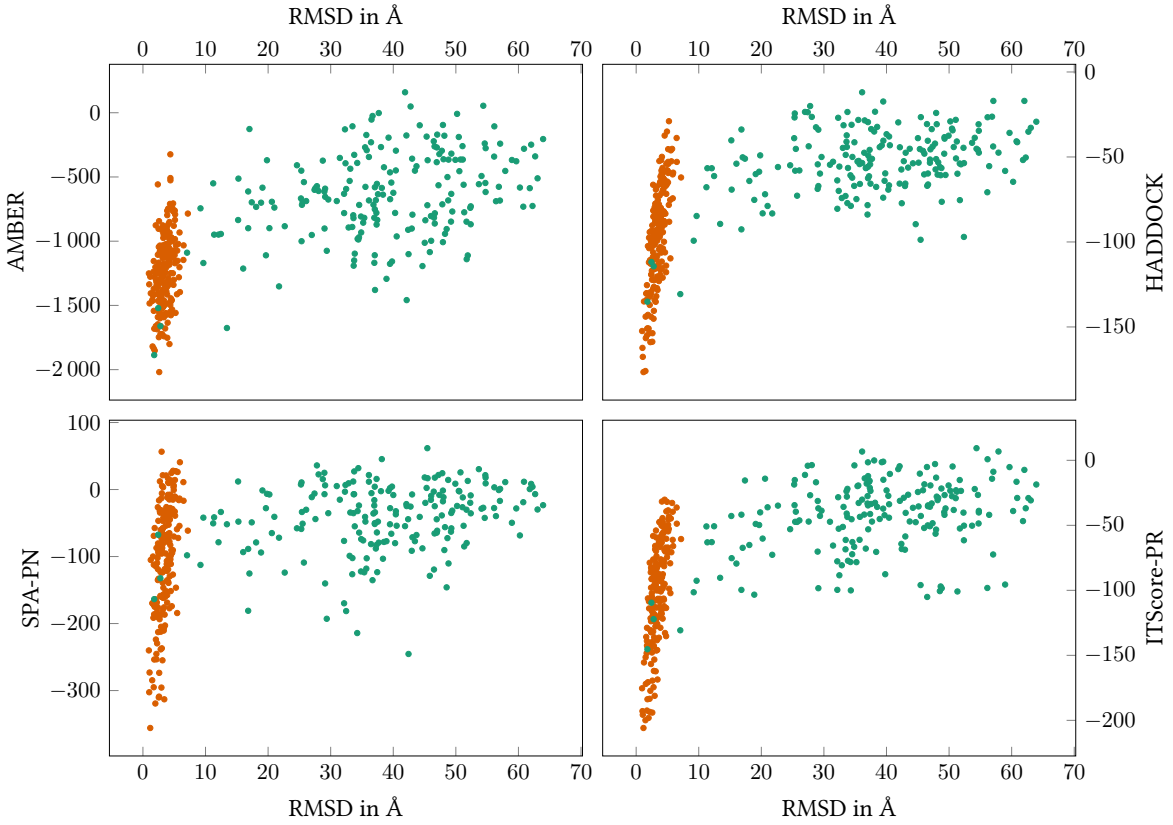


Abb. 5.13: Betrachtung der Korrelation zwischen den vier Bewertungsmodellen (AMBER, HADDOCK, SPA-PN und ITScore-PR) und der strukturellen Abweichung (RMSD) der *Decoy*-Strukturen von der Referenzstruktur 4PDB. Die *Decoy*-Strukturen sind entsprechend ihrer Herkunft (mit AIR orange, ohne AIR grün) farblich markiert.

5.4 Ergebnisse des Vergleichs

Der Vergleich der Bewertungsmodelle wurde jeweils separat an den beiden Referenzkomplexen durchgeführt und schließlich zu einer gemeinsamen Schlussfolgerung geführt. Die Reihenfolge der Referenzkomplexe wurde dabei zugunsten einer klareren Argumentation umgekehrt.

5.4.1 Referenzkomplex 4PDB

Für den Protein-RNA-Komplex 4PDB zeigten alle vier Bewertungsmodelle ein sehr ähnliches Verhalten. Wie aus der vergleichenden Abbildung 5.13 hervorgeht, schöpften die nahe-nativen *Decoy*-Strukturen annähernd den gesamten Wertebereich der jeweiligen Bewertungsfunktionen aus. Die Bewertungen der nicht-nativen Komplexe blieben hingegen auf den nicht-optimalen Bereich beschränkt, in dem die Streuung der Werte wie erwartet nicht in Korrelation zu den bestimmten RMSDs stand. Die Herkunft der Strukturen aus den beiden Grunddatensätzen spielte in dieser Beobachtung keine Rolle. Dies zeigte die Eingliederung der wenigen nahe-nativen *Decoy*-Strukturen des gänzlich zufällig erzeugten Ensembles in den mittleren bis optimalen Bewertungsbereich seines mit AIR erzeugten Counterparts. Da in allen vier Fällen die Korrelation zwischen Bewertung und RMSD im nahe-nativen Bereich gegeben und im nicht-nativen Bereich nicht erforderlich war, konnte die grundlegende Fähigkeit zur Unterscheidung zwischen nahe- und nicht-nativen Strukturen für alle betrachteten Bewertungsmodelle bestätigt werden.

Tab. 5.6: Auswahl der jeweils besten drei Strukturkandidaten für die beiden Referenzkomplexe `4PDB` und `5CMX` in Kombination mit den jeweils relevanten Bewertungsmodellen, getrennt nach Grunddatensatz. Die internen Strukturbezeichner der Dockingsimulation geben Auskunft über den Grunddatensatz (`air` mit AIR, `r` ohne Zusatzinformationen).

Referenz	Bewertungsmodell	Strukturbezeichner
4PDB	AMBER	`air94w`, `air14w`, `air74w`; `r2w`, `r47w`, `r3w`
	HADDOCK	`air7w`, `air81w`, `air25w`; `r2w`, `r5w`, `r3w`
	ITScore-PR	`air7w`, `air81w`, `air147w`; `r2w`, `r5w`, `r3w`
5CMX	ITScore-PR	`air80w`, `air19w`, `air199w`; `r51w`, `r34w`, `r95w`

Detaillierte Betrachtung des Beschreibungsverhaltens In der konkreten Ausprägung dieser Fähigkeit musste jedoch für die Bewertungsfunktionen AMBER und SPA-PN eine Einschränkung festgestellt werden. Im Vergleich zu HADDOCK und ITScore-PR war bei ihnen die Trennschärfe zwischen den beiden Klassen wesentlich geringer ausgeprägt. Dies wurde unter anderem dadurch deutlich, dass die Bewertungen einiger stark von der nativen Konformation abweichenden *Decoy*-Komplexe wesentlich stärker in den positiven Bewertungsbereich ragten als es bei den beiden anderen Vergleichsfunktionen der Fall war. Bei der Erzeugung ohne die Zusatzinformation AIR war nur mit einem sehr geringen Anteil an nahe-nativen Strukturen zu rechnen. Diese wenigen nahe-nativen Strukturen wurden jedoch durch ihre hohe Wertestreuung tendenziell ähnlich oder schlechter bewertet als einige der Ausreißer des nicht-nativen Bereichs. Eine solche Konstellation führt besonders bei mangelnder Trennschärfe wesentlich häufiger zu einer falschen Annahme über die Lokalisation und Ausformung der molekularen Schnittstelle. Dieser Effekt wurde dadurch verstärkt, dass sich das von der Dockingprozedur eingesetzte Bewertungsmaß für die Filterung der erzeugten Strukturen am Ende einer jeden Phase von der eingesetzten Bewertungsfunktion unterschied. Die energetische Bewertung mit dem Kraftfeld AMBER wies hier durch seine kompaktere Streuung, die tendenziell im optimalen Bewertungsbereich lag, noch einen deutlichen Vorteil gegenüber der Bewertung durch SPA-PN auf. Im vorliegenden Fall führten diese Umstände dazu, dass lediglich bei SPA-PN eine Konstellation entstanden ist, die sich nicht zur Auswahl nahe-nativer Vertreter aus einer vollständig zufälligen Grundmenge von *Decoy*-Strukturen eignete. Die zusätzliche Generierung weiterer *Decoy*-Strukturen würde jedoch statistischen Gesetzmäßigkeiten entsprechend durch die Ausschöpfung der genannten Streuung schrittweise zum Auffinden der nahe-nativen Konformationen führen. Auch wenn dieses Problem also durch eine starke Vergrößerung des eingesetzten *Decoy*-Ensembles umgangen werden kann, zeigt sich dabei ein wichtiger gradueller Unterschied der betrachteten Bewertungsmodelle zu Ungunsten von SPA-PN.

Überprüfung der am besten bewerteten Strukturen Im Zuge einer manuellen Verifikation wurden für beide Grunddatensätze jeweils die besten drei Strukturen entsprechend der HADDOCK-, AMBER- und ITScore-PR-Bewertung ausgewählt. Eine Übersicht dazu befindet sich in Tabelle 5.6. Die Strukturen wurden in Hinblick auf die Ähnlichkeit mit der nativen Struktur manuell untersucht. Dabei zeigte sich, dass alle unter Anwendung von AIR entstandenen Strukturen, sowie `r2w` und `r3w` eine sehr große Übereinstimmung mit der nativen Struktur aufwiesen. Dementsprechend niedrig waren die zugehörigen RMSD-Werte, die sich über einen Bereich von $1{,}17$ Å bis $2{,}79$ Å erstreckten. Die Positionierung der Komplexpartner war korrekt

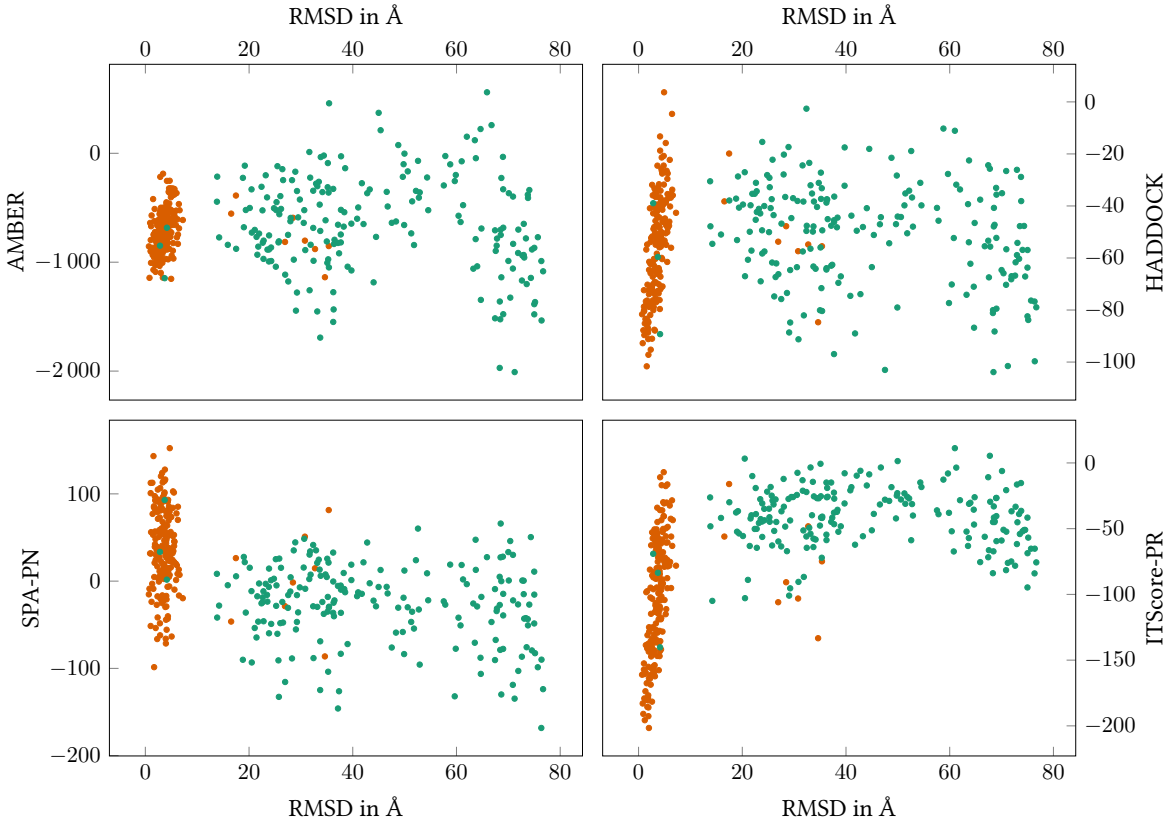


Abb. 5.14: Betrachtung der Korrelation zwischen den vier Bewertungsmodellen (AMBER, HADDOCK, SPA-PN und ITScore-PR) und der strukturellen Abweichung (RMSD) der *Decoy*-Strukturen von der Referenzstruktur `5CMX`. Die *Decoy*-Strukturen sind entsprechend ihrer Herkunft (mit AIR orange, ohne AIR grün) farblich markiert.

und wies nur geringe Rotation und Translation zur nativen Schnittstelle auf. In den *Decoy*-Komplexen `r5w` und `r47w` lag das Aptamer zwar prinzipiell an der richtigen Bindeposition und wies eine tendenziell korrekte Rotation der Längsachse auf, es befand sich jedoch jeweils in unterschiedlich ausgeprägter Schräglage entlang der molekularen Schnittstelle. Diese führte dazu, dass auf einer Seite des Aptamers stärkere Abweichungen auftraten als auf der gegenüberliegenden. Die manuelle Untersuchung ergab folglich, dass fünf der sechs betrachteten Strukturen bereits in einer hinreichend nahe-nativen Konformation vorlagen. Die verbliebene Struktur wäre trotz moderater Abweichung dazu geeignet gewesen, über die Definition von AIR bei der Suche nach nahe-nativen Strukturen zu helfen.

5.4.2 Referenzkomplex 5CMX

Im Gegensatz zum ersten Referenzkomplex `4PDB` zeichneten sich bei `5CMX` größere Unterschiede unter den vier Bewertungsmodellen ab, wie in Abbildung 5.14 zu erkennen ist. Die nahe-nativen *Decoy*-Strukturen schöpften nur bei den Bewertungen durch ITScore-PR und HADDOCK den gesamten zur Verfügung stehenden Wertebereich aus. Im Vergleich lag hier der besondere Fokus in der Aussparung der optimalen Bewertungsbereiche der beiden anderen Bewertungsmodelle. In der AMBER-Bewertung zeigte sich eine kompakte Streuung im mittleren Bereich, während die nahe-nativen Strukturen durch SPA-PN zum Großteil sogar wesentlich schlechter bewertet wurden als die meisten nicht-nativen Kompetitoren.

Flächenabhängigkeit Um das grundlegend verschiedene Bewertungsverhalten zu verstehen, müssen die Schnittstellen der beiden Referenzstrukturen 4PDB und 5CMX aus Abbildung 5.11 gemeinsam betrachtet werden. Der Referenzkomplex 4PDB deckt mit seiner molekularen Bindung entlang der Längsseite des Aptamers bereits einen relativ großen Teil der theoretisch erreichbaren Bindungsfläche ab. Die Bindung des Aptamers im Komplex 5CMX liegt hingegen am Kopf der Struktur, was zu einer vergleichsweise kleinen relativen Bindungsfläche führt. Im Gegenzug können sich jedoch bei nicht-nativen *Decoy*-Strukturen deutlich größere Bindungsflächen zwischen den beiden Komplexpartnern bilden als in der nativen Konformation. Ist das Bewertungsschema stark abhängig von der Fläche der molekularen Schnittstelle, dann besteht besonders im Falle einer kleinen nativen Schnittstelle die begründete Gefahr, dass die Bewertung von Strukturen mit größeren Bindungsflächen tendenziell verzerrt wird. Bei der energetischen Bewertung nach AMBER entsteht die Flächenabhängigkeit durch die Akkumulation der Einzelenergiewerte über die gesamte molekulare Schnittstelle, was Aspekte der Spezifität nicht berücksichtigt. Das nahezu umgekehrte Bewertungsverhalten der wissensbasierten Potentiale SPA-PN kann damit jedoch nicht vollständig erklärt werden, da die Spezifität als Optimierungskriterium ein besseres Abschneiden der Potentiale erwarten ließ. In diesem Zusammenhang wird besonders deutlich, dass das Korrelationsverhalten der Bewertung mit der Bindungsfläche ein relevantes Untersuchungskriterium darstellt. Gerade an Komplexen mit kleinen nativen Schnittstellen zeigt sich, dass die Größe der Bindungsfläche oder ein stark dazu korreliertes Bewertungsmodell kein alleiniges Kriterium für die Einschätzung der Güte einer *Decoy*-Struktur sein kann.

Detaillierte Betrachtung des Beschreibungsverhaltens Sowohl bei der HADDOCK-Bewertung als auch bei ITScore-PR zeigte sich für nahe-native Strukturen trotz der kleinen nativen Bindungsfläche des Referenzkomplexes 5CMX eine gute Korrelation zwischen Bewertung und struktureller Abweichung. Dabei wurde der Rahmen des Bewertungsmaßes bis zum Optimum ausgeschöpft. Wird jedoch der nicht-native Bereich in die Betrachtung einbezogen, so zeigten sich bei der HADDOCK-Bewertung starke Einschränkungen in der Trennschärfe zwischen nahe- und nicht-nativen Komplexstrukturen. Im unteren RMSD-Segment der nicht-nativen Strukturen war die Bewertung noch in der Lage zur Diskriminierung. Diese Fähigkeit verlor sich jedoch im oberen Bereich schnell, wie an der starken Streuung in der Bewertung erkennbar war. In einem durchmischten *Decoy*-Ensemble kann die Bewertung nach HADDOCK also nicht verlässlich eingesetzt werden. Anders verhielt es sich, wenn das betrachtete Ensemble durch geeignete Vorverarbeitung nur noch Strukturen enthielt, die der nativen Konformation bereits ähnlich waren. In diesem Fall konnte sowohl die Korrelation zwischen Bewertung und RMSD als auch die Trennschärfe im unteren RMSD-Bereich genutzt werden, um geeignete Kandidatenstrukturen zu identifizieren. Wurde der nicht-native Bereich bei der Betrachtung des Beschreibungsverhaltens der ITScore-PR in gleicher Weise einbezogen, so zeigte sich dessen Vorteil in der durchgehend vorhandenen Trennschärfe. Sie reichte an das Maß heran, welches am vorigen Referenzkomplex 4PDB gesetzt wurde, und erlaubte damit unabhängig von der Größe der Schnittstelle die Identifikation der nahe-nativen *Decoy*-Strukturen aus dem Gesamtensemble. Die wenigen nahe-nativen Vertreter aus dem ohne AIR erzeugten Teildatensatz ordneten sich ebenfalls im mittleren bis optimalen Bewertungsbereich ein. Dies zeigte, dass die Herkunft aus den beiden Grunddatensätzen keine entscheidende Rolle für die Bewertung spielte. Der großen Streuung der Bewertung im nahe-nativen Bereich führte jedoch auch zu der Empfehlung, ohne Referenzkomplex eine größere Menge an Kandidatenstrukturen im Docking zu erzeugen. Zwar

konnte im vorliegenden Fall die korrekte Struktur identifiziert werden, das Finden mehrerer Strukturen mit positiver Bewertung und ähnlich ausgeprägter Schnittstelle würde jedoch die Signifikanz des Ergebnisses erhöhen.

Überprüfung der am besten bewerteten Strukturen Im Zuge einer manuellen Verifikation wurden für beide Grunddatensätze die bereits in Tabelle 5.6 benannten besten drei Strukturen entsprechend der Bewertung ITScore-PR ausgewählt. Sie wurden in Hinblick auf die Ähnlichkeit mit der nativen Konformation manuell untersucht. Die Überprüfung erstreckte sich mangels Eignung nicht auf die verbliebenen drei Bewertungsmodelle. Die drei unter Anwendung von AIR entstandenen Strukturen zeigten eine sehr große Übereinstimmung mit der nativen Struktur, was mit den niedrigen RMSD-Werten im Bereich von $1{,}12$ Å bis $2{,}00$ Å in Übereinklang steht. Die Positionierung der Komplexpartner war durchgehend korrekt und wies nur eine geringe Rotation und Translation zum Referenzkomplex auf. Bei den drei ohne Anwendung von AIR erzeugten *Decoy*-Strukturen zeigte sich ein heterogeneres Bild. Die Struktur `r51w` wies mit einer Abweichung von $4{,}15$ Å eine etwas größere Translation und Rotation auf, bildete jedoch hinreichend genau die native Schnittstelle der beiden Bindungspartner ab, um einen Anhaltspunkt für die Ableitung von AIR zu bieten. Die beiden Strukturen `r34w` und `r95w` wiesen mit Abweichungen >10 Å zwar auf die Bindestelle der Proteinoberfläche hin, zeigten jedoch ansonsten keine Gemeinsamkeit mit der erwarteten Bindekonstellation. Die manuelle Untersuchung ergab, dass ein größeres *Decoy*-Ensemble unbedingt notwendig ist, um ohne die Zusatzinformationen der AIR nahe-native Strukturen für den Komplex `5CMX` zu finden.

5.4.3 Gewichtung der HADDOCK-Bewertung

In der bisherigen Überprüfung hat sich die HADDOCK-Bewertung mit der in Tabelle 5.3 gegebenen Standardgewichtung nach ITScore-PR als zweitbestes Bewertungsmodell herausgestellt. Der HADDOCK-Gesamtbewertung wurde auch an anderer Stelle eine gute Eignung für die Bewertung von Protein-DNA-Komplexen zugeschrieben, ohne dass jedoch dazu die einzelnen Beiträge der Kompositbewertung untersucht worden sind [388]. Es wurde daher eine Überprüfung der einzelnen Terme mit dem Ziel durchgeführt, durch die Anpassung der Standardgewichtung eine Verbesserung der Aussagekraft für die HADDOCK-Bewertung und damit eine sinnvollere Auswahl der Strukturen zwischen den Phasen des Dockings zu erreichen. Dabei konnten diejenigen Terme ausgeschlossen werden, die den Grad der Einhaltung der Distanzvorgaben AIR und anderer experimentell bestimmter Nebenbedingungen beschreiben. Zusatzinformationen dieser Art sind im Regelfall für ein Protein-Aptamer-Docking nicht verfügbar. Da die Bewertung unabhängig von der Größe der Bindefläche erfolgen sollte, fiel der zugehörige Term ebenfalls aus der Betrachtung. Auf die Darstellung der mit dieser Analyse verbundenen Abbildungen wurde aus Gründen der Übersicht verzichtet.

Korrelation der einzelnen Terme zur strukturellen Abweichung Der von HADDOCK verwendete Desolvationsterm erbrachte keinen nutzbringenden Beitrag zur Unterscheidung nahe- von nicht-nativen Decoystrukturen. Für den Referenzkomplex `5CMX` ergab sich durch die sehr ähnliche Streuung der Bewertung beider Klassen kein Informationsgewinn, während für den Referenzkomplexe `4PDB` sogar eine leicht gegensätzliche Tendenz beobachtet werden konnte. Ein genauer Blick zeigte den Grund für die geringe Eignung dieses Terms in der Bewertung der gegebenen Komplexe. Wie aus der Meldung über die Neueinführung des Desolvationsterms

Tab. 5.7: Veränderungsvorschlag für die Gewichtungsschemata der HADDOCK-Bewertungsfunktion mit den Energietermen für van-der-Waals-Interaktionen (`vdw`), Elektrostatik (`elec`), AIR (`dist`), Fläche der Schnittstelle (`bsa`) und dem Desolvationsterm (`desolv`). Siehe dazu im Vergleich Tabelle 5.3.

Dockingphase	`vdw`	`elec`	`dist`	`bsa`	`desolv`
Unsolvatisierte Phase 1	0,60	0,60	0,01	-0,01	0,20
Unsolvatisierte Phase 2	1,00	0,20	0,10	-0,01	0,10
Solvatisierte Phase	1,00	0,05	0,10	0,00	0,00

hervorging [389], wurde dieser Term ausschließlich für Protein-Protein-Interaktionen konzipiert. Ferner konnte der zugehörigen Publikation entnommen werden, dass die Trainingsmenge mit nur 24 Komplexstrukturen sehr klein war [390]. Obwohl diese Basis keine ordnungsgemäße Bewertung von Protein-Nukleinsäure-Interaktionen zulässt, wird dieser Term in der Standardkonfiguration des Dockingsystems aus Tabelle 5.3 und in den mitgelieferten Beispielkonfigurationen verwendet.

Die alleinige Betrachtung der elektrostatischen Wechselwirkungen führte in der Darstellung zu einem Bild, dass der HADDOCK-Gesamtbewertung sehr nahe kommt. Dazu zählt neben der guten Korrelation und breiten Streuung der Bewertung im nahe-nativen Bereich auch der Verlust der Trennschärfe im oberen RMSD-Sektor der nicht-nativen Strukturen bei `4PDB`. Zwar gehen die elektrostatischen Wechselwirkungen in der Vorgabe nur mit einer relativen Gewichtung von $0{,}20$ in das Gesamtergebnis der HADDOCK-Bewertung ein, ihre absoluten Werte übersteigen die der anderen Terme jedoch betragsmäßig stark. Zusammengefasst kann festgestellt werden, dass die elektrostatischen Wechselwirkungen durch ihre Größenordnung die Gesamtbeschreibung dominieren. Die einzelne Bewertung der van-der-Waals-Wechselwirkungen zeigte bei beiden Referenzkomplexen hingegen ein gutes Ergebnis. Neben der guten Korrelation im nahe-nativen Bereich umfasste dies die eingeschränkte Streuung auch im hohen nicht-nativen RMSD-Segment. Die damit erreichte Trennschärfe war zwar weniger stark ausgeprägt als bei ITScore-PR, jedoch verglichen mit der elektrostatischen Bewertung deutlich im Vorteil. Durch ihre betragsmäßig kleinen Werte wurde trotz einer Gewichtung von $1{,}0$ nur ein geringer Einfluss auf die Gesamtbewertung ausgeübt.

Modifikation der Gewichtung Die Auswertung der einzelnen Terme ergab sehr unterschiedliche Charakteristika, die es nahe legten, eine Modifikation der von HADDOCK vorgegebenen Standardgewichtung vorzunehmen. Im folgenden werden die Veränderungsvorschläge für die beteiligten Terme der HADDOCK-Bewertung kurz vorgestellt. Der Desolvationsterm wurde aufgrund seiner fehlenden Eignung in den unsolvatisierten Dockingphasen wesentlich geringer gewichtet und während der solvatisierten Phase vollständig aus der Wertung herausgenommen. Die van-der-Waals-Wechselwirkungen wurden hingegen aufgrund ihres positiven Beschreibungsverhaltens in der Wertung angehoben. Da sie in den beiden letzten Phasen bereits voll eingingen, wurde dazu nur das Gewicht für die erste Phase angepasst. Um die vorherrschende Dominanz der weniger gut geeigneten elektrostatischen Bewertung zurückzunehmen, wurde deren Gewichtung mit jeder Phase schrittweise reduziert. Selbst bei der vorgeschlagenen Gewichtung von $0{,}05$ hat dieser Term noch hinreichend Einfluss auf die Gesamtbewertung. In Tabelle 5.7 befindet sich die Übersicht über die konkreten Vorschläge für die Anpassung der Gewichtung. Da HADDOCK die interne Kompositbewertung nur zur Strukturauswahl am En-

de einer jeden Simulationsrunde einsetzt, nicht aber als Triebkraft innerhalb der eigentlichen Simulation, war der praktische Einfluss der Modifikation sehr eingeschränkt und im Gesamten eher schwach ausgeprägt. Liegt jedoch bereits bei der Auswahl der Strukturen zum Phasenübergang eine Bewertungsfunktion zugrunde, die in Bezug auf die Nativität der simulierten Komplexe eine sinnvolle Abschätzung gibt, so ist die Wahrscheinlichkeit für den Erhalt nahenativer Kandidaten am Simulationsende dadurch erhöht. Dies konnte durch die Wiederholung der Dockingsimulation unter Anwendung der modifizierten Gewichtung bestätigt werden.

5.4.4 Visualisierung der Bewertung

Die in dieser Arbeit eingesetzten geschlossen Implementierungen gaben keine Informationen über die Aufschlüsselung der Bewertungen AMBER und HADDOCK preis, sodass auf deren Basis keine Visualisierung möglich war. Die beiden wissensbasierten Paarpotentiale SPA-PN und ITScore-PR lagen hingegen als Quellcode vor. Es genügte daher ein kleiner Eingriff in den Bewertungsprozess, um über eine zusätzliche Ausgabe alle zur Visualisierung der Bewertung notwendigen Informationen bereitzustellen. Der Bewertungsalgorithmus wurde wie folgt erweitert. Für jedes Atom wurde ein Zähler eingeführt, der bei jeder relevanten Einzelpaarbewertung um den entsprechenden Teilbetrag des Atoms erhöht wurde. Zur späteren Verwertung und Visualisierung erfolgte die Persistierung dieser Atombeiträge in der Strukturdatei des überprüften *Decoy*-Komplexes, welche im Dateiformat PDB vorlag. Dieses zeilenweise aufgebaute Format spezifiziert neben einem Formatkopf und einigen Metainformationen primär die Bezeichnungen, Eigenschaften und Koordinaten der einzelnen Atome sowie deren Zuordnung zu strukturellen Einheiten. Der B-Faktor gibt über die mittlere Abweichung der Atomposition dessen intrinsische Beweglichkeit an. Da eine Bestimmung des B-Faktors nur während einer experimentellen Strukturaufklärung oder gegebenenfalls im Verlaufe einer molekulardynamischen Simulation sinnvoll möglich ist, bleibt dieses Feld bei der Dockingsimulation ungenutzt. Die atomweisen Beiträge zur Bewertung wurden daher in diesem Feld gespeichert. Gängige Werkzeuge zur molekularen Visualisierung unterstützen die Nutzung der im B-Faktor gespeicherten Information zur benutzerdefinierten Einfärbung der Struktur.

Als Beispiel einer solchen atomweisen visuellen Aufschlüsselung der Bewertung dient Abbildung 5.15. Sie zeigt die Oberflächendarstellungen der beiden auseinandergedrehten Bindepartner eines *Decoy*-Kandidaten, die mithilfe der Software PyMOL [391] entsprechend der gespeicherten Einzelbeiträge eingefärbt wurden. Neben den neutralen (weiß) Bereichen außerhalb des eigentlichen Schnittstelle zeigt die Farbgebung, dass die Beiträge innerhalb der Schnittstelle folgerichtig zum großen Teil positiv (blau) sind. Tatsächlich befinden sich diese im Intervall $(-18{,}3)$, was bei einer symmetrischen Darstellung im vollen Bereich dazu führt, dass die negativen Beiträge nicht sichtbar sind. Durch die Reduktion des Wertebereichs auf $(-4{,}4)$ wurde das Auflösungsvermögen der Abbildung zwar unterhalb von -4 stark eingeschränkt, jedoch zeigen sich nun auch Negativbeiträge. Durch die Möglichkeit der Identifizierung bindungsrelevanter Bereiche kann der Vergleich ähnlicher Strukturkandidaten wesentlich detaillierter ausgeführt werden. Im konkreten Vergleich der Struktur `Air_127w` mit den sehr ähnlichen Varianten `Air_106w` und `Air_25w` zeigte sich nach relativer Skalierung der Werte eine hohe Ähnlichkeit. Auch wenn die Größe der einzelnen Beiträge variierte, konnten ähnliche Bereiche mit positiver Auswirkung auf die Bindung festgestellt werden. Die Visualisierung lieferte daher im Rahmen der natürlichen Schwankungen konsistente Ergebnisse.

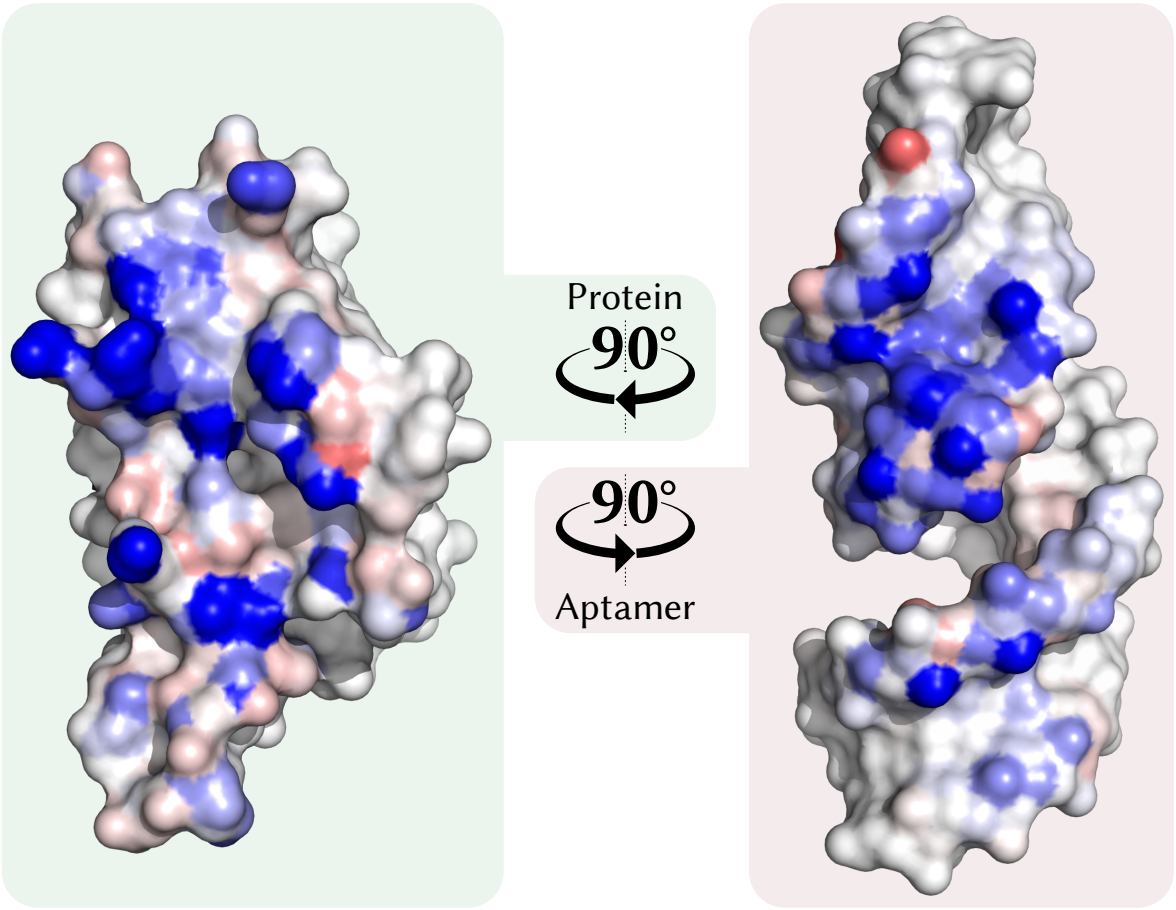


Abb. 5.15: Visualisierung der aufgeschlüsselten Bewertung nach ITScore-PR am Beispielkomplex `Air_127w` des Referenzstruktur `4PDB`. Die Teilbeträge wurden über eine Farbkodierung von blau (favorisiert) über weiß (neutral) bis rot (nicht favorisiert) auf die Oberflächen der beiden Komplexpartner aufgetragen. Im Farbbereich optimaler Bewertungen wurde zum Zweck der Erkennbarkeit eine Stauchung vorgenommen. Die beiden Bindepartner (Protein links, Aptamer rechts) wurden um die Y-Achse symmetrisch nach außen rotiert, um ihre Bindungsflächen sichtbar zu machen.

5.5 Zusammenfassung

Komplexe zwischen Proteinen und Nukleinsäuren sind nicht nur für die natürliche Vermittlung biologischer Funktionen in Organismen verantwortlich, sondern können auch durch künstlich erzeugte Bindungspartner instrumentalisiert werden. Getrieben durch atomare Wechselwirkungen entsteht dabei nicht in jedem Fall eine Schnittstelle mit größtmöglicher Bindungsfläche. Wird zur Aufklärung einer Komplexstruktur aus Protein und Nukleinsäure eine Dockingsimulation eingesetzt, dann müssen nicht nur die Tertiärstrukturen der beiden Bindepartner bekannt sein, sondern auch die simulierten Komplexe im Nachgang bewertet werden, um tatsächlich nahe-native Ergebnisse zu erhalten. Die Bewertung kann mit wissensbasierten Paarpotentialen und molekularmechanischen Ansätzen erfolgen. Über die molekularmechanische Bewertung wird die freie Energie einer solchen Bindung quantifiziert. Hier konnte gezeigt werden, dass sich eine große Bindungsfläche zwar positiv auf die energetische Bewertung auswirkt, jedoch allein noch kein Garant dafür ist, dass es sich infolge dessen um eine native oder nahe-native Bindung handelt. Während sich im Falle einer relativ großen Binderegion nur geringe Unterschiede unter den Bewertungsmodellen zeigten, konnten im Fall einer relativ kleinen Schnittstelle nur

zwei Bewertungskonzepte verlässlich zwischen nahe- und nicht-nativen *Decoy*-Strukturen unterscheiden. Durch seine hohe Trennschärfe lag ITScore-PR als bestes Modell noch vor der modifizierten internen Bewertungsfunktion der Softwaresuite HADDOCK.

Es wird daher die Empfehlung ausgesprochen, die Klassifizierung von Ergebnissen einer Dockingsimulation aus Protein und Nukleinsäure auf Basis der Bewertungsfunktion ITScore-PR vorzunehmen und gegebenenfalls die HADDOCK-Bewertungen evaluierend zu Rate zu ziehen. Ist eine Vorauswahl von nahe-nativen Strukturen getroffen, so besteht ferner die Möglichkeit, die energetische Beschreibung eines Kraftfeldes in die Analyse einzubeziehen. Diese zeigten im nahe-nativen Bereich verwertbare Korrelationen zur strukturellen Abweichung auf. Auf diese Weise kann nicht nur die energetische Beschaffenheit der Bindungsfläche beschrieben, sondern können auch Ausreißer identifiziert werden.

6 Selektion und Analyse eines Norovirus-Aptamers

Die Gruppe der Noroviren zeichnet sich durch eine hohe genetische und antigenische Diversität aus und gilt weltweit als einer der Haupterreger für epidemisch auftretende humane Virusgastroenteritis [392]. Bedingt durch das hohe Ansteckungsvermögen und die gute Beständigkeit des Virus gegenüber häufig anzutreffenden Umweltbedingungen kommt es besonders in Gemeinschaftseinrichtungen sehr schnell zu einer Ausbreitung von Norovirusinfektionen. Aus diesem Grund besteht in Deutschland für direkt nachgewiesene Fälle von Noroviruserkrankungen und teilweise bereits bei deren Verdacht eine Meldepflicht [393]. Um die oft weitreichenden und langanhaltenden Ausbrüche einzugrenzen oder gar ganz zu verhindern, sind sowohl die Detektion als auch die Verhütung von Norovirusinfektionen kritische Forschungsschwerpunkte. Neben Antikörpern bilden Aptamere eine aussichtsreiche Gruppe von Molekülen, die aufgrund ihrer biologischen Bindekapazität und der daran geknüpften Fähigkeit der gezielten Einflussnahme auf den Virus in beiden Schwerpunkten eingesetzt werden können. Im Rahmen dieser Arbeit wurde die experimentelle Selektion und bioinformatische Analyse von Aptameren durchgeführt, die in der Lage sind, am Kapsid eines Vertreters der humanen Noroviren zu binden. In den drei Vorkapiteln wurden dazu bioinformatische Analysemethoden an relevanten Beispieldatensätzen evaluiert und weiterentwickelt. Da sowohl die Anwendungsbereiche als auch die potentiell erreichbaren Analyseergebnisse jeder einzelnen Methode definierten Einschränkungen unterliegen, ist im konkreten Analysefall die Kombination der Methoden und Ergebnisse unerlässlich. Im Rahmen dieser Arbeit ist hierfür ein Verfahrensprotokoll der bioinformatischen Analyse entwickelt worden, das die Übertragung des Analyseprozesses auf Daten folgender Aptamerselektionen mit ähnlicher experimenteller Konfiguration erleichtert.

6.1 Der Norovirus als Zielstruktur der Aptamerselektion

Der erste nachgewiesene Ausbruch des Norovirus ereignete sich im Oktober des Jahres 1968 in einer Grundschule in Norwalk, Ohio, USA. Nach einem rapidem Anstieg der Infektionen innerhalb der ersten 24 Stunden infizierten sich während dieses Ausbruchs insgesamt $50\,\%$ der Lehrer und Schüler. Markant war dabei, dass die Infektion zum Teil in die Familien der infizierten weitergetragen wurde. Nachdem bakterielle Erreger ausgeschlossen wurden, begann erstmals die Suche nach einem viralen Auslöser der Erkrankung [394]. Bis zu dieser Zeit gab es noch keine bekannten Fälle akuter, viraler Gastroenteritis [395]. Schließlich gelang es im Jahre 1972, die Auslöser des Ausbruchs in Norwalk mithilfe markierter Antikörper in der Immunelektronenmikroskopie sichtbar zu machen. Sie manifestierten sich in den Aufnahmen als Partikel mit einem Durchmesser von $27\,\mathrm{nm}$ bis $32\,\mathrm{nm}$ und konnten trotz der aus heutiger Sicht niedrigen erreichten Auflösung als Viren identifiziert werden [396]. Ein Ausschnitt einer der damaligen Aufnahmen ist in Abbildung 6.1 zu sehen. Im Zuge einer späteren Klassifizierung wurde der Virus in die Familie der *Caliciviridae* eingeordnet, welche ausschließlich unbehüllte Viren mit einer einzelsträngigen, linearen RNA positiver Polarität enthält [397]. Diesem ersten Nachweis folg-

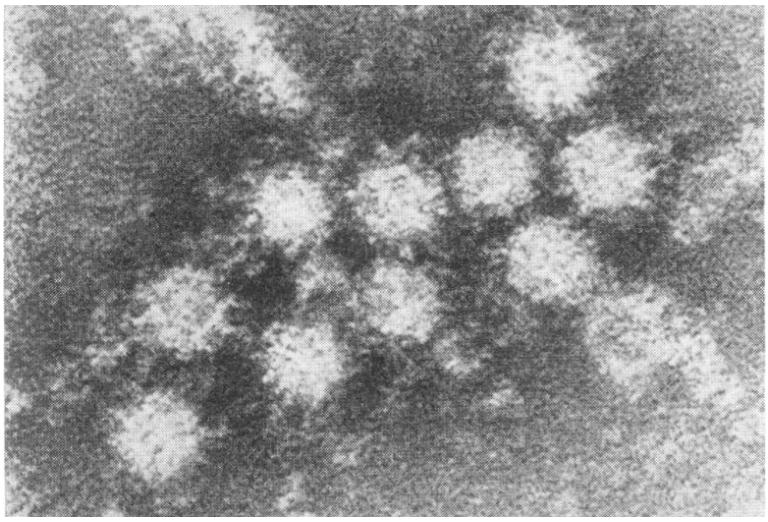


Abb. 6.1: Der Norovirus konnte erstmals durch großflächige Anlagerung markierter Antikörper experimentell sichtbar gemacht werden. Eine der ersten Aufnahmen zeigt unter Zuhilfenahme der Immunelektronenmikroskopie eine Menge von 27 nm bis 32 nm großen Partikeln in Stuhlfiltrat [396].

ten dank der steten Verbesserung der Sequenzierungs- und Analysemethoden bald auch weitere Studien, welche die weltweite Verbreitung des Norovirus in fünf verschiedenen Genogruppen belegten. Die Genogruppen `GI` und `GII`, darunter besonders der Genotyp `GII.4` [398], sind für die Mehrzahl der humanen Infektionen verantwortlich, während die anderen Genogruppen hauptsächlich bovine und murine Erreger enthalten. Ausnahme ist hier die Genogruppe `GIV`, die zwei humanpathogene Stämme enthält. Da Norovirusinfektionen selten zum Tod führen und eine zumindest kurze Immunität im Wirt hinterlassen, stehen sie unter großem Selektionsdruck. So führen Punktmutationen sowie aktive Rekombination verschiedener Norovirusstämme über Antigendrift und -shift zu der vorherrschenden großen Diversität von bereits über 150 Stämmen in 31 Genotypen [395; 399; 400]. Abbildung 6.2 gibt eine Übersicht über die Phylogenie des Norovirus.

6.1.1 Epidemiologische Aspekte

Die praktische Untersuchung der Noroviren wird seit knapp 50 Jahren durch einen wichtigen Umstand behindert [399]. Auch trotz umfangreicher Studien mit über 200 Zelllinien [402] konnten bisher keine Vertreter der humanen Noroviren reproduzierbar im Labor kultiviert werden. Ein Ansatz machte jedoch auf sich aufmerksam. Mithilfe eines 3D-Zellkulturmodells wurden Bedingungen geschaffen, unter denen ein künstliches Darmepithel heranwuchs. Zwar waren Berichten zufolge einzelne humane Norovirusstämme in diesem Gewebe zur Reproduktion in der Lage [403], die Ergebnisse dieser Studie konnten jedoch von mehreren unabhängigen Forschergruppen nicht reproduziert werden [404–406]. Ebenso konnten in verschiedenen Tiermodellen Teilerfolge erzielt werden [407–409], doch auch dieser Weg führte nicht zu einem vollumfänglichen System für die Kultivierung von humanen Noroviren [399]. Einige wichtige Erkenntnisse konnten jedoch durch Infektionsversuche mit aufgereinigten Noroviren an freiwilligen Probanden erzielt werden. Diese Versuche sind jedoch teuer, zeitintensiv und mit hohem bürokratischem Aufwand verbunden [410; 411]. Für weitere Analysen musste häufig auf kultivierbare, ähnliche Viren oder virusähnliche Partikel zurückgegriffen werden, was jedoch den Vergleich zum Originalvirus fast unmöglich macht [412; 413].

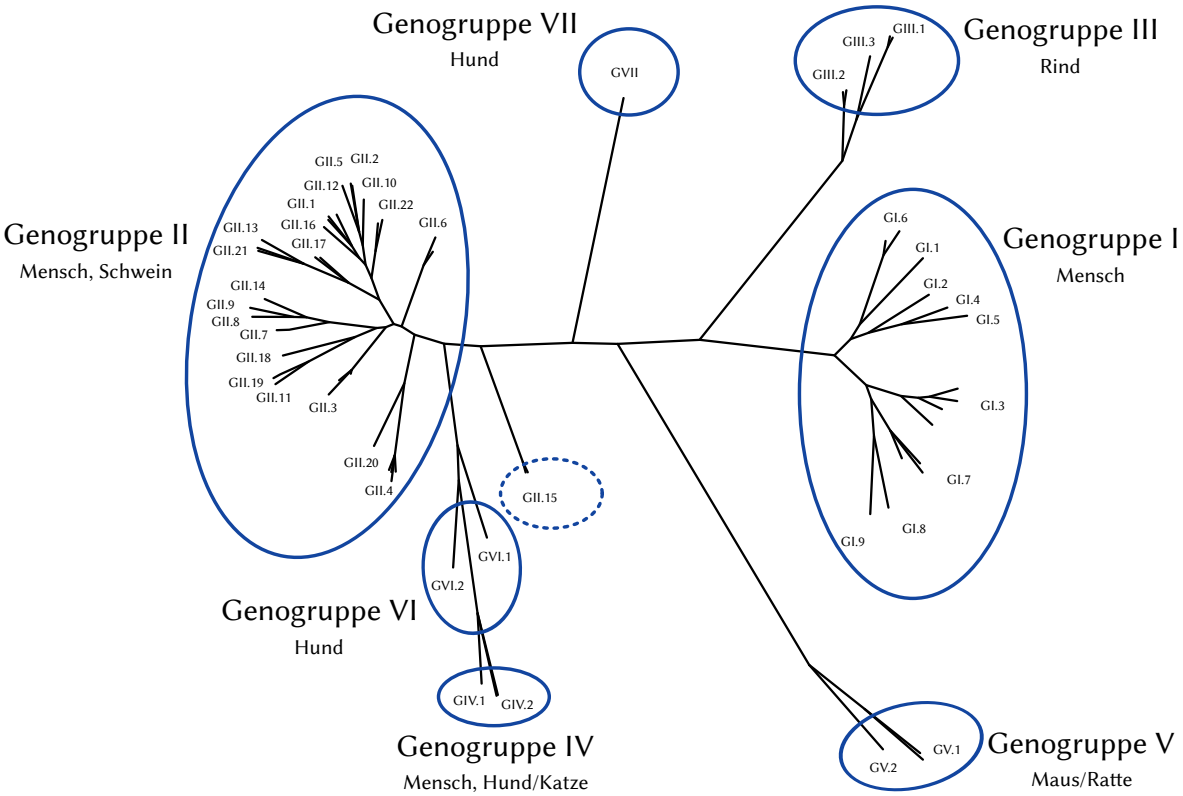


Abb. 6.2: Phylogenetische Klassifikation der Noroviren in sieben Genogruppen (GI bis GVII) basierend auf den Unterschieden der Aminosäuresequenz des kompletten VP1-Kapsidproteins [401]. Es ist besonders in den humanpathogenen Genogruppen GI und GII eine hohe Diversität zu erkennen.

Übertragung Noroviren zeichnen sich durch ihr hohes Ansteckungsvermögen aus. Die für eine Infektion notwendige Menge aufgenommener Erreger ist verglichen mit anderen Viren niedrig, auch wenn die Literatur über die genaue Quantität uneinig ist [411; 414]. Die hohe Menge an ausgeschiedener Virensubstanz [415] und die Tatsache, dass es auch nach Abklingen der Symptomatik noch weiter zur Ausscheidung infektiösen Materials kommen kann, erhöhen das Infektionspotential des Virus [416; 417]. Die Problematik wird dadurch verstärkt, dass Noroviren bei für sie günstigen Umgebungsbedingungen bis zu wochen- und monatelang ausdauern können, ohne ihr Infektionspotential zu verlieren. Ferner zeigen sich Noroviren unanfällig auf die meisten Reinigungsprodukte und sogar auf Desinfektionsmittel aus dem privaten und kommerziellen Bereich [399; 413; 418]. Es existieren zahlreiche Wege für die Übertragung des Norovirus von Mensch zu Mensch, von denen der fäkal-orale Weg und der Transport über belastete Nahrungsmittel die Hauptübertragungskanäle bilden. Ihre Rangfolge wird unterschiedlich bewertet [419–421]. Neben diesen kann der Virus jedoch auch über verunreinigtes Trinkwasser [422] oder per Schmierinfektion über kontaminierte Oberflächen [418] übertragen werden. Bei räumlicher Nähe kann eine Ansteckung darüber hinaus durch Aerosole erfolgen, die beim Erbrechen in die Umgebungsluft abgegeben werden [423]. Durch diese spezielle Kombination von Übertragungswegen kommt es gehäuft in Gemeinschaftseinrichtungen wie Krankenhäusern, Schulen und Kreuzfahrtschiffen [424–427] sowie in Gebieten mit verminderten hygienischen Bedingungen [428; 429] zu massiven Ausbrüchen.

Die Bindung an die Wirtszellen erfolgt beim Norovirus, wie häufig im viralen Umfeld zu beobachten, über spezifische Kohlehydrate, die sogenannten Kohlenhydrat-Blutgruppen-Antigene, die an der Oberfläche des intestinalen Epithels sitzen [395; 400; 430]. Nach einer Inkubations-

zeit von 20 h bis 50 h setzen die Hauptsymptome Übelkeit, Erbrechen und wässriger Durchfall ein. Begleiterscheinungen sind dabei meist mit Krämpfen verbundener Abdominalschmerz, Appetitlosigkeit und ein mit leichtem Fieber einhergehendes allgemeines Krankheitsgefühl. Da es sich um eine nichtinvasive Infektion handelt, sind Symptome wie hohes Fieber und blutiger Durchfall Ausnahmen. Während die Symptomatik in der Regel nach 8 h bis 60 h abklingt, werden noch wenigstens einige Tage lang infektiöse Partikel von den Patienten ausgeschieden [400; 411; 431; 432]. Das Auftreten von Komplikationen beschränkt sich hauptsächlich auf immungeschwächte Personen [433; 434] sowie Kleinkinder und Senioren, die schlechter mit Volumenverlust umgehen können [395]. Abhängig von der genetischen Prädestination und dem konkreten Norovirusstamm sind Teilgruppen der Bevölkerung aufgrund fehlender Rezeptoren nicht anfällig für den Virus oder nur von einer asymptomatischen Infektion betroffen [435; 436]. Nach einer Infektion ist in der Regel eine kurzzeitige Immunität von bis zu einem halben Jahr zu beobachten, die abgeschwächt auch genotypübergreifend ausgeprägt sein kann [437; 438]. Unter dem Selektionsdruck der auftretenden Herdenimmunität oder bei atypischer chronischer Infektion durch das aktive Immunsystem verändert der Virus seine Bindungsspezifität über die Mutation seiner hochvariablen P2-Subdomäne. Dies wird möglich, da die hohe Diversität der vorhandenen Kohlehydrat-Targets eine hohe Toleranz gegenüber derartigen Mutationen der viralen Binderegion erlaubt [399; 400; 439].

Gegenmaßnahmen Eine Impfung zur Norovirusimmunisierung ist derzeit nicht möglich. Es wurden jedoch mithilfe von virusähnlichen Partikeln sowohl durch die Darreichung von Nahrungsmitteln nach transgener Expression als auch durch nasale Verabreichung erste Erfolge in Form einer Immunantwort verzeichnet [440–442]. Die kurze Dauer der Immunität durch Antikörper und die hohe Diversität der Noroviren legt jedoch die Vermutung nahe, dass eine langanhaltende Immunisierung noch in weiter Ferne liegt. Es existieren einzelne Befunde über ursächliche Behandlungsmethoden, die sich jedoch klinisch bisher nicht etablieren konnten. Daher bleibt die symptomatische Behandlung der Patienten durch orale Rehydration bei gleichzeitigem Ausgleich des Elektrolytverlustes [400; 443–446].

Für den Fall eines Norovirusausbruchs sind Maßnahmen zu ergreifen, die das Ausbreiten der Infektion vermindern. Das sind zum einen die Nutzung von Schutzkleidung, also Kittel und Handschuhe, beim Umgang mit infizierten Personen und kontaminierten Gegenständen und zum anderen die geeignete Isolation bereits infizierter von gesunden Personen. Auch von offizieller Seite wird besonders zu erweiterter Handhygiene aufgerufen. Aufgrund der verminderten Wirksamkeit der häufig eingesetzten alkoholbasierten Detergenzien durch die Resistenz des Virus empfiehlt sich in diesem Fall die intensive Handwäsche mit Wasser und Seife. Für die Reinigung von Oberflächen wird hingegen eine entsprechend dosierte Lösung aus natriumhypochlorithaltiger Bleiche empfohlen [400; 447; 448]. Die Untersuchung einer Reihe von tatsächlichen Ausbrüchen zeigte, dass die eingesetzten Maßnahmen nur einen minimalen Einfluss auf die Ausbreitung der Infektion und die mittlere Dauer des Gesamtausbruchs hatten. Dabei war jedoch der recht späte Einsatzzeitpunkt der Maßnahmen fünf Tage nach Beginn des Ausbruchs auffällig [449]. Aufgrund der hohen Kontagiosität und gleichzeitig geringen Inkubationszeit ist indessen eine frühe Reaktion essentiell, damit die Maßnahmen in den Verbreitungsprozess eingreifen können, bevor er weit fortgeschritten ist.

Zudem konnte unlängst in einem Versuch gezeigt werden, dass Oberflächen aus reinem Kupfer oder hochprozentigen Kupferlegierungen sowohl die Menge aufgebrachter viraler RNA reduzieren als auch das Viruskapsid derart schädigen, dass der Erreger die Bindefähigkeit zu seinen

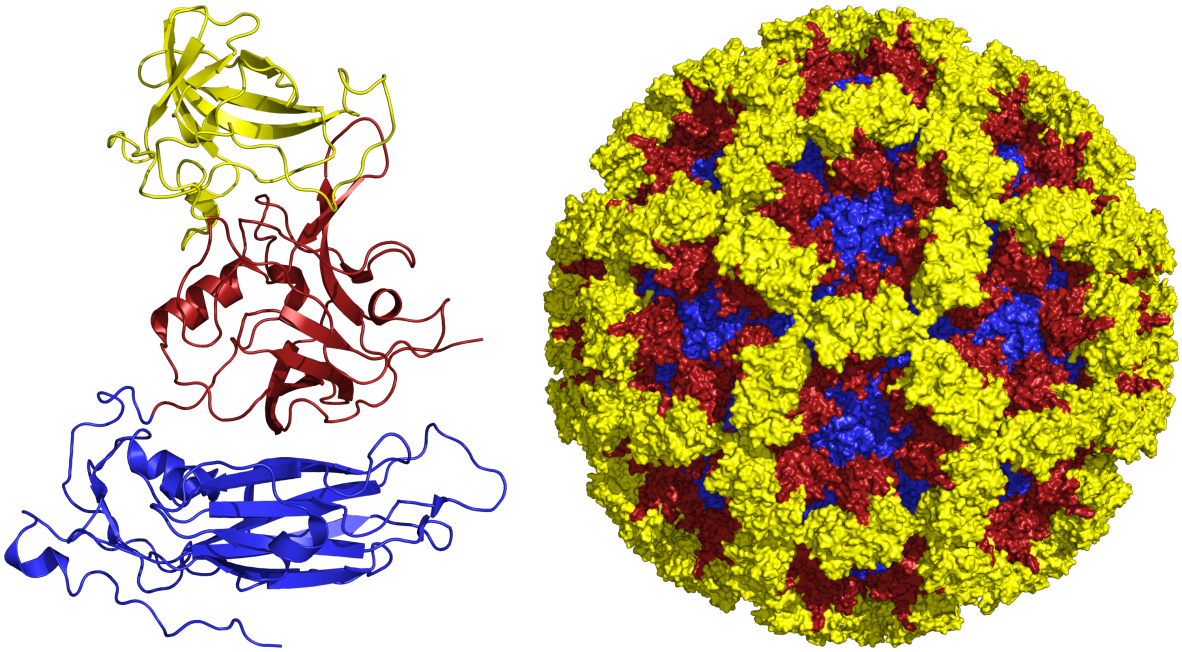
Wirtszellen zunehmend verliert. Diesen Ergebnissen zufolge eignet sich Kupfer als Oberfläche in Gemeinschaftseinrichtungen zur Verhütung der Norovirusausbreitung [450]. In einer weiteren Studie konnte zudem gezeigt werden, dass Citrat an der für die Infektion verantwortlichen Bindestelle des Noroviruskapsids binden und damit den Ausbreitungsmechanismus behindern kann. Dabei wurde beobachtet, wie sich durch die Behandlung mit Citratlösung die Oberfläche und Struktur des Kapsids derart veränderte, dass diese im Nachgang besser durch Antikörper gebunden werden konnte. Die Nutzung von citrathaltigen Reinigern stellt daher eine neue Perspektive in der Handhabung von Norovirusausbrüchen dar [451].

6.1.2 Aufbau des Norovirus

Familientypisch besitzt der Norovirus ein einzelsträngiges RNA-Genom mit positiver Polarität. Auf dem Genom mit einer Länge von $7{,}4\,\mathrm{kb}$ bis $7{,}7\,\mathrm{kb}$ befinden sich drei teilweise überlappende *Open Reading Frame*s (ORFs). Der erste Leserahmen `ORF1` kodiert ein großes Polyprotein, welches durch eine virale Protease gespalten wird. Ergebnis dieser Proteolyse sind die sechs viralen Proteine `p22`, `p48`, `VPg`, Nukleosidtriphosphatase, `3C`-ähnliche Protease und RNA-anhängige RNA-Polymerase, die jeweils große Ähnlichkeit zu den entsprechenden viralen Proteinen anderer RNA-Viren aufweisen. Die beiden Leserahmen `ORF2` und `ORF3` kodieren jeweils eins der beiden Kapsidproteine `VP1` und `VP2`, die maßgeblich für die äußere Gestalt des Virus verantwortlich sind [398; 400]. Während das größere Kapsidprotein `VP1` den Grundbaustein des Gesamtkapsids darstellt, war die Bedeutung und Lokalisierung des kleineren Kapsidproteins `VP2` lange Zeit unklar. So waren drei unabhängige, hochauflösende röntgenkristallographische Untersuchungen [452–456] nicht in der Lage, Spuren dieses Proteins zu entdecken. Erst spät konnte gezeigt werden, dass sich das `VP2`-Protein im Inneren des Virus befindet und sowohl an der Assemblierung des Kapsids als auch an der Enkapsidierung des Genoms beteiligt ist [457].

Die Struktur des Kapsidproteins VP1 Das Kapsidprotein `VP1` ist verantwortlich für die Ausbildung der Kapsidoberfläche und stellt damit den Angriffspunkt für Detektion und Inhibierung des Virus dar. Es setzt sich entsprechend der Visualisierung in Abbildung 6.3a aus zwei Hauptbestandteilen zusammen. Das ist zum einen die etwas breitere `S`-Domäne im unteren Bereich, welche im Kapsid nach innen gerichtet ist. Sie wird durch eine klassische anti-parallele β-Sandwich-Faltung charakterisiert, die häufig in viralen Kapsiden anzutreffen ist. Die beiden β-Faltblätter mit je vier beteiligten Protein-Strängen werden durch zwei α-Helices und einige *Loop*-Regionen unterbrochen. Eine dieser *Loop*-Regionen zeichnet sich unter den verschiedenen Norovirusstämmen durch eine sehr hohe Konservierung aus. Es wird vermutet, dass dieser *Loop* wichtig für die spätere Polymerbindung ist. Zum anderen befindet sich im oberen Bereich der Abbildung die eher schmale und hohe `P`-Domäne, welche im Kapsid nach Außen gerichtet ist und den Norovirus damit von anderen Viren seiner Art unterscheidet. Sie besteht aus den zwei Subdomänen `P1` und `P2`, wobei die Subdomäne `P2` sequenziell einen großen Einschub in die Subdomäne `P1` darstellt. Während `P1` durch eine β-Fass-ähnliche Struktur aus sechs β-Strängen dominiert wird, findet sich in `P2` kein solches, dominierendes Strukturmotiv wieder [452].

Die Struktur des Noroviruskapsids Das aus `VP1`-Proteinen zusammengesetzte Viruskapsid bildet eine ikosaedrige Struktur mit `T=3`-Symmetrie aus [452]. Jede der 20 gleichseitigen Dreiecksflächen eines Ikosaeders besitzt eine dreizählige Symmetrie in der Flächenmitte, drei zweizählige Symmetrien in den Zentren der angrenzenden Kanten und drei fünfzählige Symmetrien in den Eckpunkten. Durch diesen hohen Grad an Symmetrie kann eine Ikosaederstruktur aus



Monomer des Kapsidproteins VP1. **(a)** **(b)** Gesamtstruktur des Norovirus-Kapsids.

Abb. 6.3: Darstellung des Norovirus-Kapsids in Form des Grundbausteins VP1 (a) und der Gesamtstruktur (b) in einheitlichem Farbschema. Auf der geschlossenen inneren Schale, die sich aus den S-Domänen (blau) zusammensetzt, bilden sich durch die schmaleren P-Domänen (P1 in rot, P2 in gelb) deutlich erhabenen Strukturen aus [458].

genau 60 identischen asymmetrischen Einheiten aufgebaut werden. Im hier beschriebenen Fall einer T=3-Symmetrie besteht eine solche asymmetrische Einheit aus drei quasi-äquivalenten Untereinheiten [459]. Diese quasi-äquivalenten Untereinheiten werden im Noroviruskapsid durch die bekannte VP1-Struktur gebildet und sind bis auf kleine atomare Abweichungen der relativen Koordinaten und einen 19 Aminosäuren langen Präfix an einer der Untereinheiten identisch [458]. Die 180 beteiligten VP1-Proteine sind dabei in 90 Dimeren organisiert. Im Inneren des Viruskapsids entsteht somit eine geschlossene Schale aus den S-Domänen der VP1-Proteine, die in den vollen ikosaedrischen Kontakten der Struktur teilnehmen. Die schmaleren P-Domänen, welche nach oben gerichtet auf den S-Domänen sitzen, bilden eine halboffene, erhabene Struktur auf der Oberfläche des Virus aus und nehmen dabei nur an den dimerischen Kontakten teil [452]. Durch ihre exponierte Lage mit hoher Zugänglichkeit ist besonders die P2-Subdomäne mit ihrer hochvariablen Sequenzregion für die molekulare Erkennung prädestiniert. Über die genauen Mechanismen der Erkennung existieren jedoch gegensätzliche Meinungen [460; 461]. Die Kombination der beiden Domänen führt in der ikosaedrischen Gesamtkonstellation zu der in Abbildung 6.3b sichtbaren, charakteristischen Struktur des Viruskapsids mit seinen 32 kelchförmigen Vertiefungen bei einem maximalen Durchmesser von 38 nm [452]. Das Kapsid verhilft dem Norovirus nicht nur zu einer hohen thermischen Stabilität von bis zu 55 °C, sondern auch zu einer hohen Toleranz im pH-Wertbereich von 3 bis 7 [462], mit der er die natürliche Schutzfunktion der menschlichen Magensäure weitestgehend umgehen kann.

6.1.3 Nachweis des Norovirus

Bereits mit dem Beginn der ersten Norovirusausbrüche wurde aus den Beobachtungen von Symptomatik und Ausbreitungscharakteristik ein klinisch-epidemiologisches Profil erstellt, mit dessen Hilfe auch ohne Probenverfügbarkeit und Laboruntersuchung eine Einordnung vorlie-

gender Krankheitsausbrüche geschehen konnte. Die sogenannten Kaplan-Kriterien berücksichtigen sowohl den zeitlichen Verlauf als auch die Ausprägung der Symptomatik. Auch wenn auf diese Weise eine sehr hohe Spezifität und eine moderate Sensitivität erreicht werden kann [401; 463], bieten die Untersuchungen des biologischen Materials eine signifikant höhere Informationsausbeute. Die erste funktionsfähige Nachweismethode für Noroviren war die Elektronenmikroskopie, welche allein auf die visuelle Morphologie als Bewertungskriterium setzte. Da es sich dabei um eine hochgradig insensitive und teure Methode handelt, werden in heutigen Laboratorien andere Verfahren eingesetzt, die auf den folgenden zwei fundamentalen Prinzipien aufbauen. Über Enzymimmunoassays wird die Oberfläche des Virus detektiert, während die auf RT-PCR basierenden Verfahren Fragmente der viralen RNA zur Erkennung einsetzen [396; 401; 464]. Beide ermöglichen neben dem klinischen Einsatz auch die Verwendung in anderen Gebieten, wie beispielsweise in der Nahrungsmittelkontrolle.

Aktuelle Nachweismethoden Die erste Gruppe von Verfahren, die Enzymimmunoassays, nutzen speziell erzeugte Anti-Norovirus-Antikörper, die mit hoher Spezifität an der Oberfläche des Kapsids binden und dadurch ein entsprechendes Signal erzeugen. Immunochromatographische Schnelltests (*Lateral Flow Assay*) zeichnen sich dabei durch ihre geringe Anforderung an die Laborausstattung und die schnelle Durchführbarkeit in unter $15\,\mathrm{min}$ aus, weisen aber nur eine geringe Sensitivität auf, die von Genogruppe zu Genogruppe markant variiert. Eine Herausforderung ist hier speziell die hohe Diversität der Noroviren und deren anhaltender Weiterentwicklungsprozess. In handelsüblichen Kits werden aus diesem Grund meist Antikörpergemische eingesetzt, auch wenn bereits einzelne Antikörper mit einer Spezifität für ganze Genogruppen beschrieben wurden. Mit den Kits lassen sich zwar höhere Sensitivitäten erreichen als mit den benannten Schnelltests, für den sicheren Nachweis eignen sie sich jedoch trotzdem nur bedingt. Es wird daher prinzipiell geraten, die Enzymimmunoassays primär zum schnellen Screening einer großen Menge von Proben einzusetzen und anschließend die Ergebnisse durch RT-PCR zu bestätigen [401; 465–467].

Die zweite Gruppe von Verfahren detektiert nach einer Vervielfältigung mit geeigneten Primern die virale RNA. Nachdem die ersten RT-PCR-Systeme auf einer kleinen, konservierten Region des `ORF1` des Virengenoms erfolgreich waren, entwickelten sich unter der stetig wachsenden Datenlage schnell Systeme der zweiten Generation. Diese waren in der Lage, die Mehrheit der zu dieser Zeit in Umlauf befindlichen Norovirusstämme zu detektieren. Der Schritt hin zur *Reverse Transcription Quantitative PCR* (RT-qPCR) machte die bis dahin genutzten Agarose-Gele überflüssig. In den meisten Protokollen konnte die Detektion mithilfe von fluoreszenzmarkierten Oligonukleotiden nun zunehmend automatisiert werden. Die Weiterentwicklung der Verfahren resultierte in der Verringerung sowohl der falsch-negativen Ergebnisse durch eine interne Extraktionskontrolle als auch der Kreuzkontamination durch Reduktion der notwendigen experimentellen Schritte. Mit der Verfügbarkeit weiterer Sequenzdaten konnte zudem eine Region des Virusgenoms gefunden werden, die eine höhere Sensitivität bei der Detektion erlaubt. Die RT-qPCR-Systeme gelten heute als Goldstandard in der Detektion der Noroviren in Nahrungsmitteln und im klinischen Umfeld, auch wenn die Interpretation der Ergebnisse im Bereich sehr geringer Viruslast große Sorgfalt erfordert [401; 468–470].

Auftretende Probleme Die Detektion des Norovirus findet naturgemäß in sehr komplexen Stoffgemischen statt. Probematrizen wie menschlicher Stuhl und Nahrungsmittel bieten durch ihre Reichhaltigkeit an Bindeflächen und potentiellen Reagenzien im experimentellen Ablauf ein

großes Risiko der Inhibition und Kreuzreaktion. Kontrollproben können hier nur eine gewisse Abhilfe schaffen. Besondere Bedeutung bekommt diese Problematik bei der ohnehin schwierigen Interpretation von Ergebnissen der RT-qPCR bei sehr niedriger Virenlast, da sich die benannten Störeffekte in diesem Fall deutlich stärker auswirken [399; 400]. Zu beachten ist auch, dass sich die Detektion in beiden beschriebenen Methoden auf charakteristische Teile des Virus beschränkt, welche allein nicht zwingend das tatsächliche Vorhandensein infektiöser Partikel folgern lassen. So könnten durchgeführte Enzymimmunoassays auch auf leere Viruskapside oder auf nach Degradation verbliebene Teilstücke dieser ansprechen. Ebenso besteht die Möglichkeit, dass die relativ kurzen Zielsequenzen der RT-qPCR-Systeme in teilweise degradierter Virus-RNA detektiert werden. Diese kann nicht nur in beschädigten Viruspartikeln, sondern auch frei in der Matrix vorliegen. Das hiervon verursachte Potential falsch-positiver Ergebnisse sollte bei der Auswertung stärker in den Blick genommen werden. Allgemein werden die Schwierigkeiten bei der Interpretation der Ergebnisse trotz steigender Nutzung der Verfahren bisher kaum thematisiert [399].

Einsatz von Aptameren Bedingt durch ihre Fähigkeit, an eine Vielzahl möglicher Zielmoleküle zu binden [48], bilden die Nukleotidaptamere eine alternative Klasse von Erkennungselementen für den Einsatz in modifizierten Immunoassays [471]. Sie können analog zu Antikörpern spezifisch für die Oberfläche des Noroviruskapsids selektiert und anschließend zur Detektion eingesetzt werden. Aptamere bieten gegenüber den bisher sehr häufig eingesetzten Antikörpern einige Vorteile. Sie zeichnen sich durch eine höhere Stabilität und biologische Zugänglichkeit aus, können aber ähnliche Affinitäten wie Antikörper vorweisen. Sowohl die *in vitro* stattfindende Produktion als auch die spätere chemische Synthese bieten Vorteile bezogen auf Produktionszeit, -komplexität und -kosten [472]. Durch die Möglichkeit der Ankopplung weiterer molekularer Bestandteile an die Aptamere reicht ihre Nutzung über die reine Detektion hinaus. Zusätzlich könnte durch die Aptamerbindung in geeigneter Konstellation die Interaktionsfläche des Viruskapsids für die Kohlenhydrat-Blutgruppen-Antigene blockiert und damit sein Reproduktionszyklus behindert werden. Aptamere besitzen damit ein hohes Potential für die Diagnose und Behandlung von Norovirusinfektionen.

6.1.4 Eingesetzter Norovirusstamm

Unter den verschiedenen humanen und teilweise-humanen Genogruppen des Norovirus ist die Genogruppe `GII` global die dominierende. Auch unterhalb dieser Gruppe hat sich ein Genotyp als dominant den anderen gegenüber herausgestellt. So werden Schätzungen zufolge $80\,\%$ aller Norovirusausbrüche dem Genotyp `GII.4` zugeschrieben [473; 474]. Ein Grund dafür ist die große genetische Variabilität, die den Vertretern der Genogruppe `GII` im Vergleich zu denen der Gruppe `GI` ein effektiveres Adaptionsverhalten bei auftretender Immunität verleiht. Die Durchsetzungsfähigkeit des Genotyps `GII.4` hängt zudem eng mit der Entwicklungsgeschichte des verwandten Genotyps `GII.3` zusammen. Diese beiden sind sich hinreichend ähnlich, sodass `GII.3`-spezifische Antikörper ebenfalls Kreuzreaktionen mit den Viren des Genotyps `GII.4` aufwiesen. Einige schwere Ausbrüche des Genotyps `GII.3` in den späten 1990er Jahren führten daher beiläufig auch zu einem gewissen Grad an Immunität gegenüber dem Genotyp `GII.4`, was als zusätzlicher Selektionsdruck treibend in dessen Entwicklung wirkte. Der Effekt der schnelleren Entwicklung von Virenstämmen unter Einfluss höherer Wirtsimmunität konnte bereits beim Influenzavirus nachgewiesen werden [475–477]. So entstehen auch innerhalb der Genogruppe `GII.4` in regelmäßigen Abständen von wenigen Jahren jeweils neue, genetisch un-

Tab. 6.1: Sequenz des VP1-Kapsidproteins des Norovirus Genotyp GII.4 der Clustergruppe Farmington Hills.

1	MKMAS NDANP SDGST ANLVP EVNNE VMALE PVVGA AIAAP VAGQQ NVIDP	50
51	WIRNN FVQAP GGEFT VSPRN APGEI LWSAP LGPDL NPYLS HLARM YNGYA	100
101	GGFEV QVILA GNAFT AGKII FAAVP PNFPT EGLSP SQVTM FPHII VDVRQ	150
151	LEPVL IPLPD VRNNF YHYNQ SNDPT IKLIA MLYTP LRANN AGEDV FTVSC	200
201	RVLTR PSPDF DFIFL VPPTV ESRTK PFTVP ILTVE EMTNS RFPIP LEKLF	250
251	TGPSG AFVVQ PQNGR CTTDG VLLGT TQLSP VNICT FRGDV THIAG THNYT	300
301	MNLAS QNWNN YDPTE EIPAP LGTPD FVGRI QGMLT QTTRG DGSTR GHKAT	350
351	VSTGD VHFTP KLGSI QFNTD TNNDF ETGQN TRFTP VGVVQ DGNGT HQNEP	400
401	QQWVL PSYSG RTGHN VHLAP AVAPT FPGEQ LLFFR STMPG CSGYP NMNLD	450
451	CLLPQ EWVQH FYQEA APAQS DVALL RFVNP DTGRV LFECK LHKSG YVTVA	500
501	HTGQH DLVIP PNGYF RFDSW VNQFY TLAPM GNGTG RRRAL LE	542

terscheidbare Varianten des Virus, die sich schnell verbreiten [474]. Eine dieser Varianten wurde erstmals im Jahr 2002 bei einem Ausbruch in Farmington Hills, Michigan, USA identifiziert und verbreitete sich in mehreren Staaten der USA [426].

In der vorliegenden Studie wurde das große Kapsidprotein VP1 des Norovirus Genotyp GII.4 der Clustergruppe Farmington Hills als Zielstruktur für die Aptamerselektion eingesetzt. Die Einordnung der Genogruppe in die gesamte Phylogenie des Norovirus ist Abbildung 6.2 zu entnehmen. Die Aminosäuresequenz des genutzten Proteins ist in Tabelle 6.1 gegeben.

6.2 Experimentelle Durchführung und Auswertung

Ein Ziel dieser Arbeit ist die Selektion und Analyse eines DNA-Aptamers, welches spezifisch an das große Kapsidprotein VP1 des Norovirus Genotyp GII.4 der Clustergruppe Farmington Hills bindet. Alle im Labor durchgeführten Experimente dieses Abschnitts wurden durch Mitarbeiter der TU Dresden geplant und durchgeführt. In den jeweiligen Unterabschnitten befinden sich weiterführende Hinweise zur Zuordnung der Verantwortlichkeiten. Für die Selektion wurde das vielfach bewährte Screeningverfahren SELEX eingesetzt. Basierend auf einer zufällig chemisch synthetisierten Oligonukleotidbibliothek setzt das Verfahren quasi keine Informationen über das eingesetzte Zielmolekül voraus. Die experimentelle Anordnung limitiert jedoch die Größe der Bibliothek, sodass nur ein kleiner, zufälliger Bruchteil des bei gegebener Sequenzlänge theoretisch möglichen Sequenz- und Strukturraums in der Bibliothek repräsentiert wird. Es waren komplexe bioinformatische Analysen notwendig, um trotz der Einschränkungen einen möglichst großen informationellen Gewinn zu schaffen. Zur Bildung einer dafür geeigneten Datenbasis waren konventionelle Sequenzierungstechnologien wie beispielsweise die Sanger-Sequenzierung [478] nicht ausreichend. Neben dem zeitintensiven Klonierungsprozess, der hierfür notwendig ist, stellt besonders der recht niedrige Grad der sequenziellen Abdeckung einen kritischen Engpass dar. In dieser Arbeit wurde daher die relativ neue Technologie NGS verwendet. Sie erlaubt die kostengünstige Erfassung sehr großer Mengen von Sequenzdaten in relativ kurzer Zeit [479] und eignet sich daher sogar für die Sequenzierung der Bibliotheken

Tab. 6.2: Sequenztemplate der initialen Oligonukleotidbibliothek, mit der Modifikation `G` → `C` an Position 5 des Vorwärtsprimers übernommen aus [480].

5′	GCCTCTTGTGAGCCTCCTAAC - $\mathbf{N_{49}}$ - CATGCTTATTCTTGTCTCCC	3′

mehrerer Selektionsrunden. Eine erste Validierung der erfolgreichen Selektion geschah sowohl über die Beobachtung der Anreicherung von Aptamerkandidaten innerhalb der Bibliothek als auch über ein konkretes Bindeexperiment.

6.2.1 Aptamerselektion

Nach der Vorbereitung der Oligonukleotidbibliothek und der Zielproteine wurde die Selektion in zwölf Positiv- und drei darin eingebetteten Negativselektionsrunden durchgeführt und die entsprechenden Bibliotheken wurden sequenziert. Als Abweichung vom Standardablauf kommt den Negativselektionsrunden eine besondere Bedeutung zu. Sie sind notwendig, um die Spezifität der Aptamere zu erhöhen, da sich die Zielmoleküle im späteren Einsatzfeld in einer komplexen Matrix mit zahlreichen unterschiedlichen Bindeflächen befinden. Die Planung und Durchführung der Aptamerselektion erfolge durch die Mitarbeiter der Bioverfahrenstechnik des Instituts für Naturstofftechnik der TU Dresden.

Vorbereitung Die initiale Oligonukleotidbibliothek des SELEX-Laufes wurde aus $49\,\mathrm{nt}$ langen, zufälligen Insertsequenzen gebildet, die beidseitig von notwendigen Primern flankiert wurden. Der sequenzielle Aufbau wurde dabei aus einem bereits vorher genutzten Template übernommen [480]. Eine Modifikation `G` → `C` an Position 5 des Vorwärtsprimers vermied den Effekt der Primerhomodimerbildung und resultierte im finalen Template in Tabelle 6.2. Die Synthese der Bibliothek erfolgte durch die Firma Ella Biotech GmbH, Martinsried, Deutschland. Nach einer initialen Denaturierung für die Dauer von $10\,\mathrm{min}$ bei einer Temperatur von $90\,^{\circ}\mathrm{C}$ erfolgte eine Abkühlung der Oligonukleotidbibliothek für weitere $15\,\mathrm{min}$ auf Eis. Die langsame Angleichung auf Raumtemperatur erfolgte anschließend ohne Hilfsmittel. Die direkte Folge aus De- und Renaturierung führte zu einer Neufaltung der Oligonukleotide, die eventuell vorhandene strukturelle Abweichungen und Deformationen löste. Das als Zielmolekül der Selektion gewählte `VP1`-Kapsidprotein des Norovirus Genotyp `GII.4` wurde durch die Firma Ribbox GmbH, Radebeul, Deutschland als rekombinantes Protein exprimiert. Zur späteren Immobilisierung wurde das Zielprotein am `C`-Terminus mit einem Hexahistidin-Tag markiert.

Selektion In der ersten der zwölf positiven Selektionsrunden wurden $2\,\mathrm{nmol}$ der Zielproteine eingebracht, in den folgenden elf Runden jeweils $0{,}2\,\mathrm{nmol}$. Das Zielprotein wurde dabei im SELEX-Bindepuffer (siehe Tabelle 6.3) mit der Oligonukleotidlösung der Bibliothek inkubiert. Nach einer Behandlung in einem Rotationsmischer bei Raumtemperatur für $60\,\mathrm{min}$ wurde die Mischung auf die *His SpinTrap*-Säule der Firma GE Healthcare, Pollards Wood, Großbritannien geladen. Diese Einweg-Anionenaustauscher-Säule erlaubt die Aufreinigung Polyhistidin-getaggter Proteine mit dem schonenden und effektiven Verfahren der immobilisierten Metallaffinitätschromatographie. Anschließend wurde die Säule mit dem Waschpuffer (siehe Tabelle 6.3) gereinigt. Die verbliebenen, gebundenen DNA-Protein-Komplexe wurden schließlich mithilfe des Elutionspuffers (siehe Tabelle 6.3) von der Säule entfernt. Die einzelnen Chemikalien, aus denen sich die Puffer, wie in Tabelle 6.3 vorgestellt, zusammensetzen, wurden von der Firma

Tab. 6.3: Übersicht über die genaue Zusammensetzung sowie die pH-Einstellung der im Experiment verwendeten Pufferlösungen. Die einzelnen Chemikalien, aus denen sich die Puffer zusammensetzen, wurden von der Firma Merck KGaA, Darmstadt, Deutschland, bezogen.

Pufferlösung	**pH**	**Zusammensetzung der Lösung**	
SELEX-Bindepuffer	6	100 mM NaCl	17 mM Na_2HPO_4
		3 mM KH_2PO_4	5 mM KCl
		2 mM $MgCl_2$	1 mM $CaCl_2$
		0,02 $\%_{Vol}$ TWEEN® 20	
Waschpuffer	6	100 mM NaCl	17 mM Na_2HPO_4
		3 mM KH_2PO_4	5 mM KCl
		2 mM $MgCl_2$	1 mM $CaCl_2$
		0,2 $\%_{Vol}$ BSA	0,02 $\%_{Vol}$ TWEEN® 20
Elutionspuffer	4	100 mM NaCl	7 mM Na_2HPO_4
		5 mM KCl	2 mM $MgCl_2$
		1 mM $CaCl_2$	0,02 $\%_{Vol}$ TWEEN® 20
		pH-Anpassung mit Zitronensäure	

Merck KGaA, Darmstadt, Deutschland bezogen. Nach jeweils drei dieser Positivselektionsrunden wurde eine zusätzliche Negativselektion durchgeführt. Mit dieser Strategie konnten Oligonukleotide aus der Bibliothek entfernt werden, welche an die experimentell erforderlichen Hintergrundmaterialien oder die spätere Probenmatrix, also menschlichen Stuhl, banden. Zu diesem Zweck wurde eine Stuhlsuspension auf die Säule aufgetragen, von der sichergestellt war, dass sie keine Noroviruskapsidproteine enthält. Nach mehreren Reinigungsschritten mit dem SELEX-Bindepuffer (siehe Tabelle 6.3) wurde die Oligonukleotidlösung ebenfalls auf die Säule gegeben. Oligonukleotide, die an der so vorbereiteten Säule gebunden haben, wurden wegen ihrer unspezifischen Bindung aus der Bibliothek entfernt. Zwischen den jeweiligen Runden wurden die verbliebenen Oligonukleotide gereinigt und amplifiziert, um eine ausreichend große Bibliothek für die darauffolgende Runde zu erhalten. Das PCR-Protokoll schloss 20 Zyklen mit einer Anlagerungstemperatur von 58 °C und einer Elongationszeit von jeweils 30 s ein.

Sequenzierung Während aufgrund der limitierten Kapazitäten der herkömmlichen Sequenzierungsverfahren nur eine Auswahl der Sequenzen der letzten Selektionsrunde erfassbar gewesen wäre, konnten mit NGS selbst intermediäre Runden in hoher Abdeckung erfasst und der Auswertung zur Verfügung gestellt werden. In Vorbereitung auf die Sequenzierung wurden die Oligonukleotide durch das Hinzufügen einiger zusätzlicher Sequenzabschnitte erweitert. Dies umfasste sowohl eine Schnittstelle für die Kopplung an die Durchflusszelle des Sequenzierers als auch für die Durchführung der Sequenzierung notwendige Primer. Zur späteren Zuordnung der gelesenen Sequenzen zu den untersuchten Bibliotheken wurden zusätzliche Indexprimer und spezifische Barcode-Sequenzen verwendet. Nach dieser Vorbereitung konnten die Proben, die zwischen den jeweiligen Positivselektionen entnommen wurden, durch die *Deep Sequencing Group* des *Biotechnology Center* der TU Dresden sequenziert werden. Der NGS-kompatible Sequenzierer HiSeq2000 der Firma Illumina arbeitet nach dem Prinzip des *Sequencing by Synthesis* (SBS) und generierte dabei 76 nt lange Einzelreads.

Tab. 6.4: Ergebnisse der Sequenzierung. Für jede Positivselektionsrunde ist der interne Rundenbezeichner, die Konzentration nach der 1. PCR sowie die Anzahl der eingelesenen Sequenzen gegeben.

#	Name	Konzentration	Sequenzen	#	Name	Konzentration	Sequenzen
1	FH-6-01	$10{,}6\,\mathrm{ng\,\mu L^{-1}}$	374 780	7	FH-6-07	$10{,}0\,\mathrm{ng\,\mu L^{-1}}$	347 643
2	FH-6-02	$19{,}2\,\mathrm{ng\,\mu L^{-1}}$	54 417	8	FH-6-08	$8{,}9\,\mathrm{ng\,\mu L^{-1}}$	376 367
3	FH-6-03	$13{,}4\,\mathrm{ng\,\mu L^{-1}}$	112 001	9	FH-6-09	$8{,}6\,\mathrm{ng\,\mu L^{-1}}$	430 755
4	FH-6-04	$8{,}5\,\mathrm{ng\,\mu L^{-1}}$	117 809	10	FH-6-10	$10{,}4\,\mathrm{ng\,\mu L^{-1}}$	441 842
5	FH-6-05	$16{,}4\,\mathrm{ng\,\mu L^{-1}}$	293 958	11	FH-6-11	$9{,}2\,\mathrm{ng\,\mu L^{-1}}$	320 939
6	FH-6-06	$7{,}7\,\mathrm{ng\,\mu L^{-1}}$	415 885	12	FH-6-12	$10{,}0\,\mathrm{ng\,\mu L^{-1}}$	348 892

Die Ergebnisse dieser Sequenzierung finden sich als Übersicht in Tabelle 6.4. Die experimentellen Bedingungen führten dazu, dass die Sequenzausbeute für die Runden 2 bis 4 deutlich niedriger ausgefallen ist als für die restlichen Runden. Mit über $50\,000$ Sequenzen wurde aber auch hier eine nutzbare Grundlage für eine Evaluation der Anreicherung geschaffen. Da sich die weiteren bioinformatischen Analysen ausschließlich auf die letzten Runden bezogen, deren Abdeckung mit über $300\,000$ Sequenzen sehr gut war, wurde die Sequenzierung als erfolgreich angesehen.

6.2.2 Evaluation der Anreicherung von Aptamerkandidaten

Über die Beobachtung der Bibliotheksentwicklung konnte ein tieferes Verständnis des SELEX-Prozesses gewonnen werden. Im Ergebnis lieferte die Sequenzierung für jede der zwölf Positivselektionsrunden eine Datei, welche die gelesenen Sequenzen im FASTQ-Dateiformat enthielt. Dieses erfasst neben der sequenziellen Rohform eine nukleobasenweise Annotation mit einem in Buchstabenform kodierten Qualitätswert, der eine grobe Aussage über die Wahrscheinlichkeit von Sequenzierungsfehlern gibt. Nach einer initialen Nachbearbeitung der bereitgestellten Rohdaten wurde die Anreicherung von Aptamerkandidaten in der Bibliothek für jede Runde über zwei numerische Diversitätsmaße bestimmt. Der Verlauf dieser Kennwerte wurde anschließend verifiziert.

Nachbearbeitung der Sequenzdaten Obwohl die Reads mit einer Länge von $76\,\mathrm{nt}$ hardwarebedingt kürzer waren als die $90\,\mathrm{nt}$ langen, primerflankierten Templatesequenzen, enthielten sie in den meisten Fällen die vollständigen Insertsequenzen. Kleine Verschiebungen im Leseraster machten es für die Freistellung der Insertsequenzen erforderlich, die genauen Positionen der flankierenden Primersequenzen zu bestimmen. Aufgrund ihrer unvollständigen Erfassung während der Sequenzierung musste an dieser Stelle neben der expliziten Lesefehlertoleranz auch die Verkürzung der Primersequenzen berücksichtigt werden. Im Zuge einer halbautomatischen Sichtung des Sequenzmaterials ergaben sich zwei Lesefehler für den Vorwärtsprimer und drei Fehler für den Rückwärtsprimer als sinnvolle Schwellwerte für die Erkennung. Fehlerhafte Konstellationen der Primer führten während der Erkennungsprozedur zu einem Ausschluss der Sequenz aus dem Datensatz. Diese umfassten neben fehlenden auch zu kurze und damit nicht signifikante Primersequenzen sowie Fälle, in denen die detektierte Insertsequenz von der erwarteten Länge abwich.

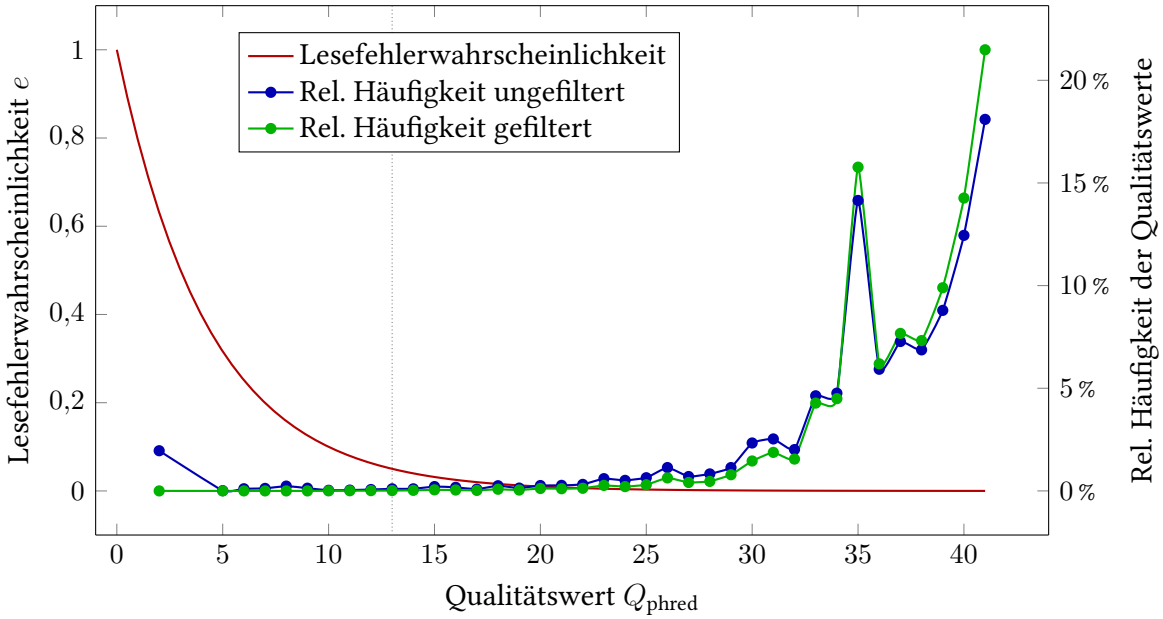


Abb. 6.4: Bedeutung des Qualitätswertes im FASTQ-Format. Bereits ab einem Qualitätswert von 13 (gepunktete Linie) beträgt die Lesefehlerwahrscheinlichkeit (rote Linie) für die bewertete Nukleobase $\leq 5\,\%$. An den relativen Häufigkeiten, mit denen die Qualitätswerte in den sequenzierten Daten vor (blau) und nach (grün) der Anwendung des Qualitätsfilters auftraten, sind die hohe Qualität der erhaltenen Sequenzen und der Effekt der Filterung gut zu erkennen.

Der nukleobasenweise notierte Qualitätswert hat nach seiner Dekodierung einen Wertebereich von 0 (sehr wahrscheinlich ein Lesefehler) bis 41 (sehr wahrscheinlich kein Lesefehler), welcher nach dem Prinzip der Phred-Bewertung zu interpretieren ist. Der Qualitätskennwert leitet sich entsprechend Formel 6.1 von der approximierten Wahrscheinlichkeit e für einen Lesefehler beim Sequenzieren einer Nukleobase ab, wobei sich das in Abbildung 6.4 (rote Linie) gezeigte, logarithmische Verhalten ausbildet. Trotz der allgemein hervorragenden Sequenzierungsqualität kam es zu einzelnen wahrscheinlich fehlerhaft bestimmten Nukleobasen, die sich in den Qualitätskennwerten in Abbildung 6.4 niederschlagen. Generell wird in diesem Fall eine Filterung der Sequenzen anhand der nukleobasenweisen Qualitätsbewertung vorgeschlagen [481]. Dem entsprechend wurde ein Filter entwickelt, der eine Sequenz genau dann als potentiell fehlerhaft einstuft, wenn einer der folgenden beiden Grenzwerte unterschritten wird. Der strenge Grenzwert regelt dabei mit einem Qualitätswert von 10 den Ausschluss einzelner Nukleobasen mit sehr niedriger Bewertung. Der weiche Grenzwert hingegen bezieht sich auf zusammenhängende Bereiche und die Gesamtsequenz. Er erlaubt das maximale Auftreten von vier einzelnen oder zwei zusammenhängenden Nukleobasen mit einem Qualitätswert kleiner als 20. Diese Filterkriterien sind zwar im Gesamten sehr restriktiv, entfernten jedoch aufgrund der guten Sequenzierungsqualität nur einen kleinen Teil der Sequenzen aus dem Datensatz.

$$Q_{\text{phred}} = -10 \cdot \log_{10} e \tag{6.1}$$

Die statistische Verteilung der Qualitätswerte offenbarte, dass die erreichte Sequenzierungsqualität am Anfang der Reads im Durchschnitt am höchsten war und zur Readmitte hin nahezu linear, jedoch nur sehr leicht abfiel. In der hinteren Hälfte des Reads war ein etwas stärkerer Abfall zu beobachten (nicht gezeigt). Das beobachtete Qualitätsprofil entlang der Reads trat hardwarebedingt auf und führte nicht zu einer Einschränkung der Nutzbarkeit der Sequenzdaten.

Wie an der Häufigkeitsverteilung der einzelnen Qualitätswerte in Abbildung 6.4 erkennbar ist, lag das Qualitätsprofil durchgängig in einem hohen Qualitätsbereich. In den ungefilterten Daten lagen die Qualitätskennwerte im Mittel von 30 bis 38, nach Anwendung des Filters im Mittel von 33 bis 40. Durch die Filterschritte wurden dabei aus dem Datensatz rundenunabhängig etwa $2\,\%$ der Sequenzen aufgrund nicht identifizierbarer Primersequenzen und etwa $19\,\%$ wegen Verstoßes gegen die Qualitätskriterien entfernt. Da es sich bei den mitgelieferten Qualitätskennwerten des Sequenzierers lediglich um Schätzungen handelt, können einzelne fehlerhaft sequenzierte Nukleotide unerkannt bleiben. Ebenso kann es im Einzelfall innerhalb der PCR zu ungewollten Punktmutationen kommen. Im Rahmen der Evaluation der Anreicherung waren diese im Mittel konstanten Störgrößen vernachlässigbar. Für die geplanten bioinformatischen Analysen wurde jedoch ein weiterer Filterschritt angefügt, der sehr selten vorkommende Sequenzvertreter aus dem Datensatz entfernte. Nach mehreren Selektionsrunden war die Wahrscheinlichkeit sehr hoch, dass die einzeln vorkommenden Oligonukleotide als Fehlerprodukt irrelevant für die weiteren Analysen waren. Für die letzte Sequenzierungsrunde ergibt sich daher ein Datensatz von $251\,444$ gültigen Einzelsequenzen, die in 5703 eindeutigen Ausprägungen vorlagen.

Validierung der Anreicherung auf Basis der Diversität Die Anreicherung von Aptamerkandidaten in einer SELEX-Bibliothek kann ähnlich zur Entwicklung einer Population mehrerer Spezies betrachtet werden. Jede der vorherig genannten Ausprägungen steht dabei für eine Spezies und umfasst eine definierte Anzahl von Einzelsequenzen als Individuen. Im vorliegenden Fall wurde diese Analogie zugrunde gelegt. In der direkten Folge konnte über bekannte Maße die Diversität der Bibliothek bestimmt und von ihr numerisch verwertbar auf die Anreicherung geschlossen werden. Formal wurde die sequenzierte Bibliothek dazu in S unterschiedliche Spezies unterteilt. Auf Basis der Zuordnung der Einzelsequenzen wurden anschließend die relativen Häufigkeiten p_i ihres Vorkommens in der Bibliothek bestimmt. Für die Bewertung der Diversität wurden schließlich die Shannon-Entropie H (Formel 6.2) sowie der modifizierte Simpson-Index D' (Formel 6.3) genutzt [482]. Trotz großer Gemeinsamkeit in der generell adressierten Charakteristik, zeigen diese beiden Indizes einen markanten Unterschied. Generell stützt die Shannon-Entropie ihre Bewertung auf den Grad der Ungewissheit bei zufälligem Ziehen aus der Population. In einem hoch angereichertem Sequenzpool gibt die Anreicherung dabei bereits im Voraus eine klare Tendenz vor, welche Sequenz bei einem zufälligen Zug mit hoher Wahrscheinlichkeit erhalten wird. Entsprechend ist auch die Bewertung der Diversität durch die Shannon-Entropie gering. Dies führt zu einer leichten Abnahme der Gewichtung von häufig vorkommenden zugunsten von seltenen Spezies. Der originale Simpson-Index beschreibt die Diversität einer Population als die Wahrscheinlichkeit, beim Ziehen zweier zufälliger Individuen die gleiche Spezies zu erhalten. Als Wahrscheinlichkeit nimmt der Simpson-Index daher Werte von 0 bis 1 an. Da die einzelnen relativen Häufigkeiten quadratisch in die Berechnung eingehen, werden seltene Spezies im Simpson-Index schwächer gewichtet und niedrige Diversitäten schlecht innerhalb des Wertebereiches aufgelöst. Die logarithmische Transformation des modifizierten Simpson-Indexes erhöht seine Auflösung im niedrigen Bereich und verändert seinen Wertebereich, der nun beginnend mit 0 nach oben begrenzt ist.

$$H = -\sum_{i=1}^{S} p_i \cdot \ln p_i \tag{6.2}$$

$$D' = -\ln\left(\sum_{i=1}^{S} {p_i}^2\right) \tag{6.3}$$

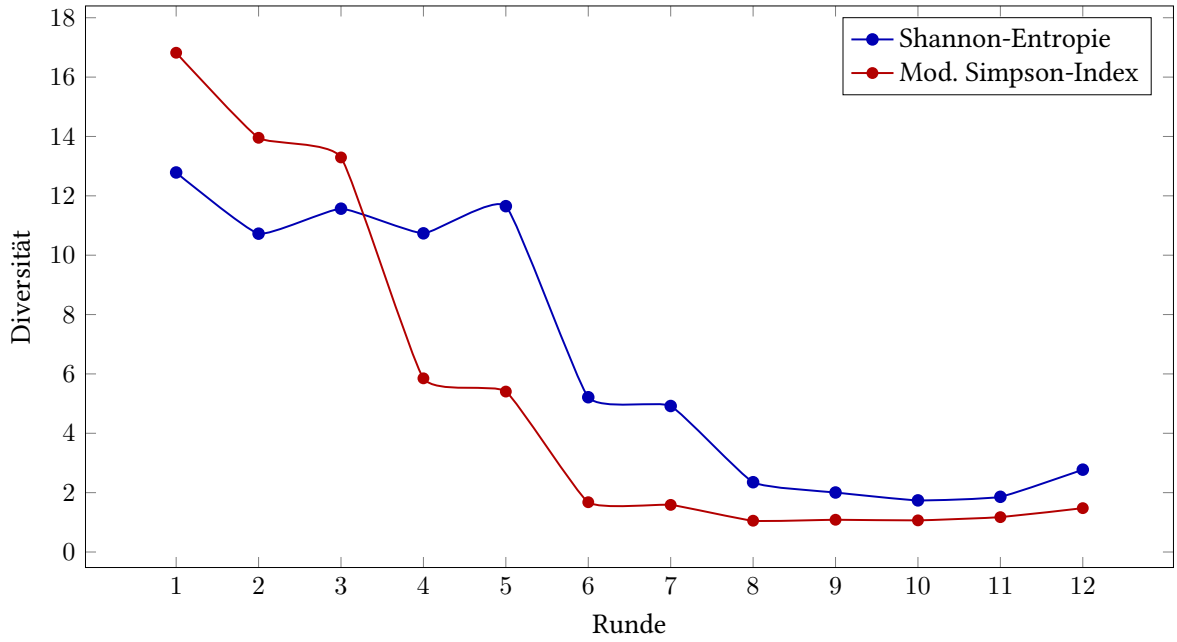


Abb. 6.5: Die Diversität wurde sowohl mit der Shannon-Entropie (blau) als auch mit dem modifizierten Simpson-Index (rot) nach allen zwölf Positivselektionsrunden bestimmt. Die Maße zeigen im Detail ein abweichendes Verhalten, welches jedoch in beiden Maßen einem schubweisen Abwärtstrend folgt. Dieser deutet auf eine tatsächliche Anreicherung der Aptamerkandidaten während des Selektionsprozesses hin, die jedoch nach Runde 8 stagniert.

Beide Diversitätsmaße zeigen mit den in Abbildung 6.5 ersichtlichen individuellen Charakteristiken eine Verringerung der Diversität über die Runden des SELEX-Prozesses an, die auf eine erfolgreiche Anreicherung von tatsächlich bindenden Aptamersequenzen schließen lässt. Aufgrund mangelnder Daten konnte der Effekt der ersten Selektionsrunde nicht bemessen werden. Über diese hinaus zeigte der modifizierte Simpson-Index den deutlich klareren Trend. Neben kleineren Schwankungen der Diversität im üblichen Bereich einer experimentellen Durchführung und Sequenzierung zeigten sich drei markante Veränderungen der Anreicherung. Die erste Verringerung der Diversität von Runde 1 zu Runde 2 war eher leicht ausgeprägt und ließ sich bei manueller Inspektion der Sequenzdaten nicht nachempfinden. Auch wenn sie sich bei beiden Diversitätsmaßen gleichermaßen manifestierte, handelte es sich bei ihr offenbar um einen ungewollten Nebeneffekt der schlechten Sequenzausbeute in der zweiten Sequenzierungsrunde. Die Anreicherung, die mit der Diversitätsverringerung von Runde 3 zu Runde 4 korrespondierte, umfasste nur eine kleine Menge an Sequenzen, deren Häufigkeit entsprechend stark zugenommen hatte. Da häufig vorkommende Sequenzen im Kontext der Shannon-Entropie geringer gewichtet werden als solche mit niedrigen Häufigkeiten, sprach die Shannon-Entropie nur sehr verhalten auf diese Anreicherung an. Als Grund für den heftigen Einbruch der Diversität kann mit großer Sicherheit die erste Negativselektionsrunde festgestellt werden, die zwischen den Runden 3 und 4 stattfand. Durch die erstmalige Applikation der komplexen Matrix aus menschlichem Stuhl wurden zahlreiche Oligonukleotide aus der Bibliothek entfernt, was während der 4. Runde die Konkurrenzsituation innerhalb der Bibliothek markant veränderte. Die Anreicherung zwischen den Runden 5 und 6 zeichnete sich in der Shannon-Entropie stärker ab als im modifizierten Simpson-Index. Das lässt darauf schließen, dass in diesem Fall eine größere Anzahl von Sequenzen jeweils schwächer von der Anreicherung betroffen war. Eine manuelle Inspektion der Bibliothek konnte diese Folgerung bestätigten. Alle weiteren Veränderungen waren eher geringfügig. Da

sich während eines SELEX-Experiments nahezu keine neuen Sequenzen durch Mutation herausbilden, war bereits in der zweiten Negativselektionsrunde zwischen Runden 6 und 7 kein Effekt mehr zu erkennen. Insgesamt stagnierte die Anreicherung den Kennwerten zufolge ab Runde 9. Aus den Kennwerten kann also dank der neuen Sequenzierungsmethode nachempfunden werden, dass während der Selektion tatsächlich eine Anreicherung von möglichen Aptamerkandidaten in der Bibliothek stattgefunden hat. In der detaillierten Cluster-Analyse in Abschnitt 6.3 erfolgt eine genauere Beleuchtung der Anreicherung mit grafischer Visualisierung.

6.2.3 Experimentelle Verifikation der Bindung

Für ein vielversprechendes Aptamer aus der letzten Sequenzierungsrunde wurde die tatsächliche Aptamer-Target-Bindung mittels *Surface plasmon resonance* (SPR)-Spektroskopie experimentell verifiziert. Oberflächenplasmonen sind evaneszente Elektronendichteschwankungen, die parallel zur Oberfläche eines Metalls verlaufen und deren Intensität exponentiell zur Ausbreitungslänge abnimmt. Die physikalische Grundlage für dieses Phänomen ist, dass sich die delokalisierten Leitungselektronen in einem Metall ihren Eigenschaften nach als Plasma betrachten lassen. Oberflächenplasmonen lassen sich nur unter bestimmten Bedingungen durch eine ebenfalls evaneszente Welle anregen, die beispielsweise durch die Totalreflexion von Licht an der Außenfläche eines Prismas entsteht. Bei der SPR-Spektroskopie wird der Effekt ausgenutzt, dass die Wellenlänge der Oberflächenplasmonen sehr empfindlich auf Änderungen im Brechungsindex reagiert, die in der unmittelbaren Nähe der Metalloberfläche stattfinden. Mit der Wellenlänge der Oberflächenplasmonen ändert sich auch die von ihnen aus dem Lichtstrahl ausgekoppelte Energiemenge, sodass ein charakteristisches Minimum in der Lichtreflexion messbar wird. Da die Adsorption biologischer Moleküle auf einer speziell vorbereiteten Membran zu einer Änderung ihres Brechungsindexes führt, kann die Bindung in Echtzeit mithilfe der SPR-Spektroskopie gemessen werden [483–485]. Die Planung und Durchführung der experimentellen Verifikation der Aptamerbindung erfolge durch die Mitarbeiter der Bioverfahrenstechnik des Instituts für Naturstofftechnik der TU Dresden.

Experimentelle Durchführung Im Experiment wurde das ^{li}SPR-System der Firma *capitalis technology* GmbH, Berlin, Deutschland eingesetzt. Das System besteht aus einem Spektrometer mit dazugehöriger Software, einem Probenhandlingsystem, einer On-Chip-Mikrofluidik und den entsprechenden Sensorchips. Zur Anregung der Oberflächenplasmonen wurde kollimiertes LED-Licht der konstanten Wellenlänge von $810\,\mathrm{nm}$ eingesetzt. Nach der Reflexion durch die Goldfläche und erneuter Kollimierung erfolgte schließlich die Messung durch eine CCD-Kamera [486]. Für den Laufpuffer (siehe Tabelle 6.5) wurde eine Flussrate von $5\,\mu\mathrm{L\,s^{-1}}$ und eine Temperatur von $30\,^{\circ}\mathrm{C}$ festgelegt. Vor jedem experimentellen Schritt wurde der Chip in einem dreistufigen Verfahren gereinigt. Zuerst wurde die Goldschicht mit 10 Tropfen $65\,\%$-er rauchender Salpetersäure gespült, anschließend wurde der Chip für $2\,\mathrm{min}$ in Neutralisierungslösung (siehe Tabelle 6.5) inkubiert und ausgiebig mit bidestilliertem Wasser ausgespült. Die in einer Konzentration von $0{,}5\,\mathrm{mg\,mL^{-1}}$ in Wasser aufgelösten Noroviruskapsidproteine wurden mit der freien Goldoberfläche für $1\,\mathrm{h}$ bei Raumtemperatur inkubiert. Der nun mit Protein überzogene Sensorchip wurde nochmals gründlich mit dem Laufpuffer (siehe Tabelle 6.5) gespült. Für die experimentelle Überprüfung der Bindefähigkeit wurde der häufigste in der Bibliothek der letzten Selektionsrunde vorkommende Aptamerkandidat entsprechend Tabelle 6.6 ausgewählt. Um sein Bindungsverhalten mit dem Zielprotein zu bewerten, wurde $2\,\mathrm{min}$ nach Beginn

Tab. 6.5: Übersicht über die genaue Zusammensetzung sowie die pH-Einstellung der Pufferlösungen, die während der experimentellen Verifikation der Aptamer-Target-Bindung verwendet wurden. Die einzelnen Chemikalien, aus denen sich die Puffer zusammensetzten, wurden von der Firma Merck KGaA, Darmstadt, Deutschland, bezogen.

Pufferlösung	**pH**	**Zusammensetzung der Lösung**	
Laufpuffer	6	$100\,\mathrm{mM}$ NaCl	$17\,\mathrm{mM}$ Na_2HPO_4
		$3\,\mathrm{mM}$ KH_2PO_4	$5\,\mathrm{mM}$ KCl
		$2\,\mathrm{mM}$ $MgCl_2$	$1\,\mathrm{mM}$ $CaCl_2$
		$0{,}02\,\%_{\mathrm{Vol}}$ TWEEN® 20	
Neutralisierungslösung		$1 \times 25\,\%$ wässrige Ammoniaklösung (NH_3)	
		$1 \times 30\,\%$ Wasserstoffperoxid (H_2O_2)	
		$5 \times$ bidestilliertes Wasser	

Tab. 6.6: Sequenz des mittels SPR-Spektroskopie verifizierten DNA-Aptamers, welches nach der letzten Selektionsrunde am häufigsten in der Bibliothek zu finden war.

$5'$	GTCTGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGGC	$3'$

des Experiments eine Lösung mit $10\,\mu\mathrm{M}$ des ausgewählten Aptamerklones für $10\,\mathrm{min}$ auf die Proteinoberfläche des Chips gegeben. Zur Beobachtung des Dissoziationsverhaltens wurde die Probenlösung im Nachgang für etwa $3\,\mathrm{min}$ durch den Laufpuffer (siehe Tabelle 6.5) ersetzt. Aus dem charakteristischen Minimum des reflektierten Lichtes berechnete das experimentelle System über die gesamte Dauer des Vorgangs ein numerisches Signal, welches für die weitere Auswertung aufgezeichnet wurde. Als Kontrollprobe wurde die Polymerase eines murinen Norovirusstammes ausgewählt, da diese dem Kapsidprotein des humanen Stammes sowohl im isoelektrischen Punkt pI ($6{,}4$ / $5{,}8$) als auch im molekularen Gewicht ($59{,}4\,\mathrm{kDa}$ / $59{,}8\,\mathrm{kDa}$) sehr ähnlich war. Der experimentelle Ablauf wurde in der gleichen Konfiguration mit dieser Kontrollprobe wiederholt.

Ergebnis der Verifikation Im Sensogramm der Abbildung 6.6 zeigt sich ein charakteristisches Bild. Beim Hinzufügen des untersuchten Aptamers stieg das Signal durch die beginnende Bindung zwischen Aptamer und Zielprotein exponentiell an. In der darauffolgenden Äquilibriumsphase stabilisierte sich das Signal mit vernachlässigbaren, unsystematischen Schwankungen. Nach Hinzugabe des Laufpuffers zur finalen Reinigung und Dissoziation zeigte sich eine schwache Verringerung des Signals, da sich unter diesen Bedingungen nur ein kleiner Teil der gebundenen Aptamere löste. Ein vom zeitlichen Ablauf ähnliches, aber von der Stärke seiner Ausprägung sehr viel schwächeres Verhalten zeigte sich bei der Durchführung mit der Negativprobe. Da das Sensogramm ein typisches exponentielles Verhalten ohne Erreichen des Sättigungsbereiches und ohne große Sprünge und Spitzen vorwies, kann von einer guten experimentellen Qualität ausgegangen werden [487]. Im Vergleich von Assoziations- und Dissoziationsphase zeigt sich die stabile Bindung zwischen Aptamer und Zielprotein. Zwar war aufgrund unspezifischer Interaktionen auch eine geringe Anlagerung des Aptamers an der Kontrollprobe zu beobachten, es war jedoch in der Lage, zwischen seinem Zielprotein und der Kontrollprobe zu unterscheiden.

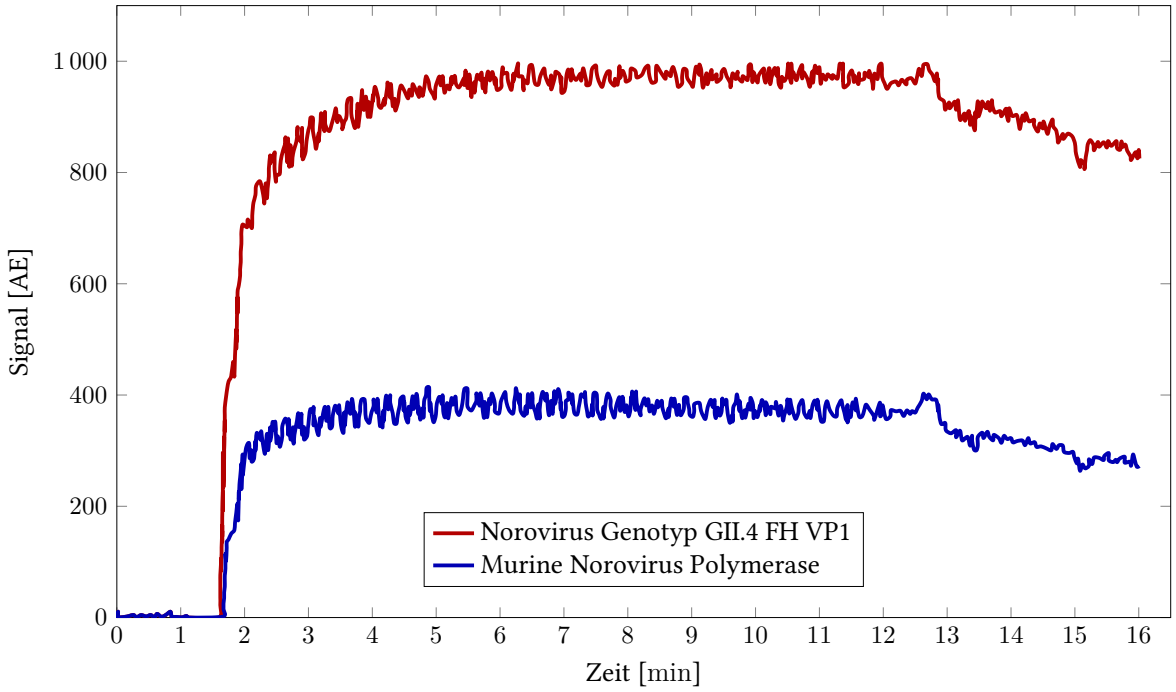


Abb. 6.6: Die Abbildung zeigt zwei Sensogramme, die mittels SPR-Spektroskopie aufgenommen wurden. Diese zeigen die Anlagerung des überprüften Aptamerkandidaten an das Zielprotein (rot) und eine Kontrollprobe (blau, enthält ein Protein, das dem Zielprotein ähnlich ist). Die Zugabe der Aptamere 2 min nach Experimentbeginn ging mit einem unmittelbaren, typisch exponentiellen Anstieg des Signals einher. Dieser war von einer mehr-minütigen Äquilibriumsphase und anschließend schwachen Dissoziation gefolgt. Für die Kontrollprobe ergab sich ein sehr schwacher Signalverlauf. Die Beobachtungen weisen auf die spezifische und stabile Bindung des Aptamers mit dem Zielprotein hin.

6.2.4 Entwurf eines bioinformatischen Analyseprotokolls

Nachdem in diesem Unterkapitel sowohl die Anreicherung innerhalb der Bibliothek als auch die Bindefähigkeit des ausgewählten Aptamers bestätigt werden konnte, galt die experimentelle Phase der Aptamerselektion als erfolgreich abgeschlossen. Die Kombination aus der zufälligen Natur der Oligonukleotidbibliothek und den experimentellen Limitierungen führte jedoch dazu, dass die gefundenen Aptamere lediglich als optimale Lösung eines kleinen Teilbereiches des theoretisch möglichen Sequenz- und Strukturraums angesehen werden konnten. Dies beeinträchtigte zwar nicht grundlegend den Einsatz des gefundenen Aptamers, gab jedoch Grund zu einer weiterführenden Untersuchung. Durch die hohe Abdeckung der neuen Sequenzierungstechnologie eröffneten sich interessante Möglichkeiten der bioinformatischen Analyse.

Möglichkeiten der bioinformatischen Analyse Neben der sequenziellen Primärstruktur kann bei derartigen Analysen sowohl auf die Sekundär- als auch auf die Tertiärstrukturinformationen der Aptamere und des Zielproteins zurückgegriffen werden. In den drei vorherigen Kapiteln wurden aus diesem Grund bereits unterschiedliche bioinformatische Analysemethoden an relevanten Beispieldatensätzen evaluiert und weiterentwickelt. In den Betrachtungen der drei Kapitel wurde deutlich, wie stark sich die Anwendungsbereiche und die analytischen Möglichkeiten der einzelnen Methoden unterscheiden. Die hohe Verfügbarkeit von Sequenzdaten auf Seite der Aptamere führt in Verbindung mit der guten algorithmischen Erschließung der Sekundärstrukturvorhersage zu zahlreichen Möglichkeiten der statistischen Auswertung. Hier sei neben der Korrelation von sequenziellen und strukturellen Merkmalen mit den Affini-

täten der jeweiligen Aptamerkandidaten auch die Detektion signifikant gehäuft vorkommender Motive in den Daten erwähnt. Die Möglichkeit, paarweise Interaktionen zwischen Aptamer und Zielprotein auf atomarer oder makromolekularer Ebene aus den vorliegenden Sequenzdaten abzuleiten, besteht jedoch nicht. Dreidimensionale Daten zu den Tertiärstrukturen und tatsächlichen Bindungskonstellationenen von Aptamer und Zielprotein können diese Informationen bereitstellen, sind jedoch wesentlich schwerer verfügbar. Sowohl der Weg über die experimentelle Strukturaufklärung als auch der über die Simulation sind mit großem Aufwand verbunden, sodass in der Regel nur sehr wenige solcher Strukturen bereitstehen. Gegeben dieser Einschränkungen ist die Kombination der Analyseverfahren für die elaborierte Gesamtanalyse von besonderer Wichtigkeit. Sie wurde daher in diesem Abschnitt weiter konkretisiert und in Form eines Verfahrensprotokolls festgehalten. Die Durchführung des Protokolls erfolgt in den folgenden zwei Abschnitten dieses Kapitels.

Beschreibung des Protokolls Die Aminosäuresequenz des Zielproteins bildet gemeinsam mit den Sequenzdaten der Aptamere aus der letzten Runde des SELEX-Experiments die Grundlage für das weitere bioinformatische Analyseverfahren, welches in zwei Hauptpfaden abläuft. Während der gesamten Analyse werden die Teilergebnisse beider Pfade miteinander abgeglichen, um die Schwächen der einzelnen Verfahren durch Ergänzung abzumildern oder auszugleichen. Eine übersichtliche Visualisierung des Verfahrensprotokolls befindet sich in Form eines Workflows in Abbildung 6.7.

Der erste Hauptpfad verfolgt die Ableitung von Sequenz- und Strukturmotiven aus den vorhandenen Sequenzdaten, die für die Bindung zwischen Aptamer und Zielprotein relevant sind. Hier ergänzen sich Verfahren, die einerseits mit biologischem Referenzkriterium und andererseits ohne ein solches auf statistisch-symbolischer Ebene arbeiten. Die für diese Betrachtung erforderliche, hohe Grundmenge an Sequenzen kann durch das Sequenzierungsverfahren zwar für die Aptamerbibliothek gewonnen werden, steht jedoch für das Zielprotein nicht zur Verfügung, sodass sich dieser Pfad auf die Auswertung der Aptamere beschränkt. In einem ersten Prozessschritt werden die Sekundärstrukturen der Aptamere anhand der verfügbaren Sequenzen bestimmt. Aktuelle Softwarelösungen zur Bestimmung der Sekundärstrukturen sind zeitlich sehr effizient und erlauben daher auch bei großen Sequenzpools eine lückenlose strukturelle Abdeckung. Über die optionale Ausgabe suboptimaler Strukturensembles besteht ferner die Möglichkeit, in der Analyse feingliedrig auf strukturelle Variabilität einzugehen. Eine initiale Clusteranalyse gibt Auskunft über das Vorhandensein von Subgruppierungen, welche aufgrund unterschiedlicher sequenzieller Merkmale im weiteren Verlauf gegebenenfalls einer gesonderten Behandlung bedürfen. Über die Boltzmann-Relation kann aus den Auftretenshäufigkeiten der einzelnen Aptamerkandidaten eine Schätzgröße ihrer relativen Affinitäten abgeleitet werden, da eine Ausgangsbibliothek ohne sequenziellen Bias verwendet wurde. Gemeinsam mit dieser und einer Beschreibung der Aptamere durch numerischen Deskriptoren sind über Regressionsverfahren Rückschlüsse darauf möglich, welche Nukleotide positiv beziehungsweise negativ an der Zusammensetzung der Gesamtaffinität des Komplexes aus Aptamer und Zielprotein beteiligt sind. Als Gegenpol zum hohen Detailgrad der abgeleiteten Informationen stehen sowohl die algorithmische Beschränkung auf einen Teil der Sequenzen als auch das relativ starre Konzept der Abbildung von Sequenzmotiven über Deskriptoren. Hier kommen die Vorteile von Mustersuchalgorithmen zur Geltung, die das Einfließen aller Sequenzinformationen erlauben und sequenzielle Variabilität unterstützen. Im Vergleich der Ergebnisse zeigt sich bereits auf dieser inneren Ebene der Verfahrensstruktur der positive Effekt der Ergänzung. Die vorgestellten

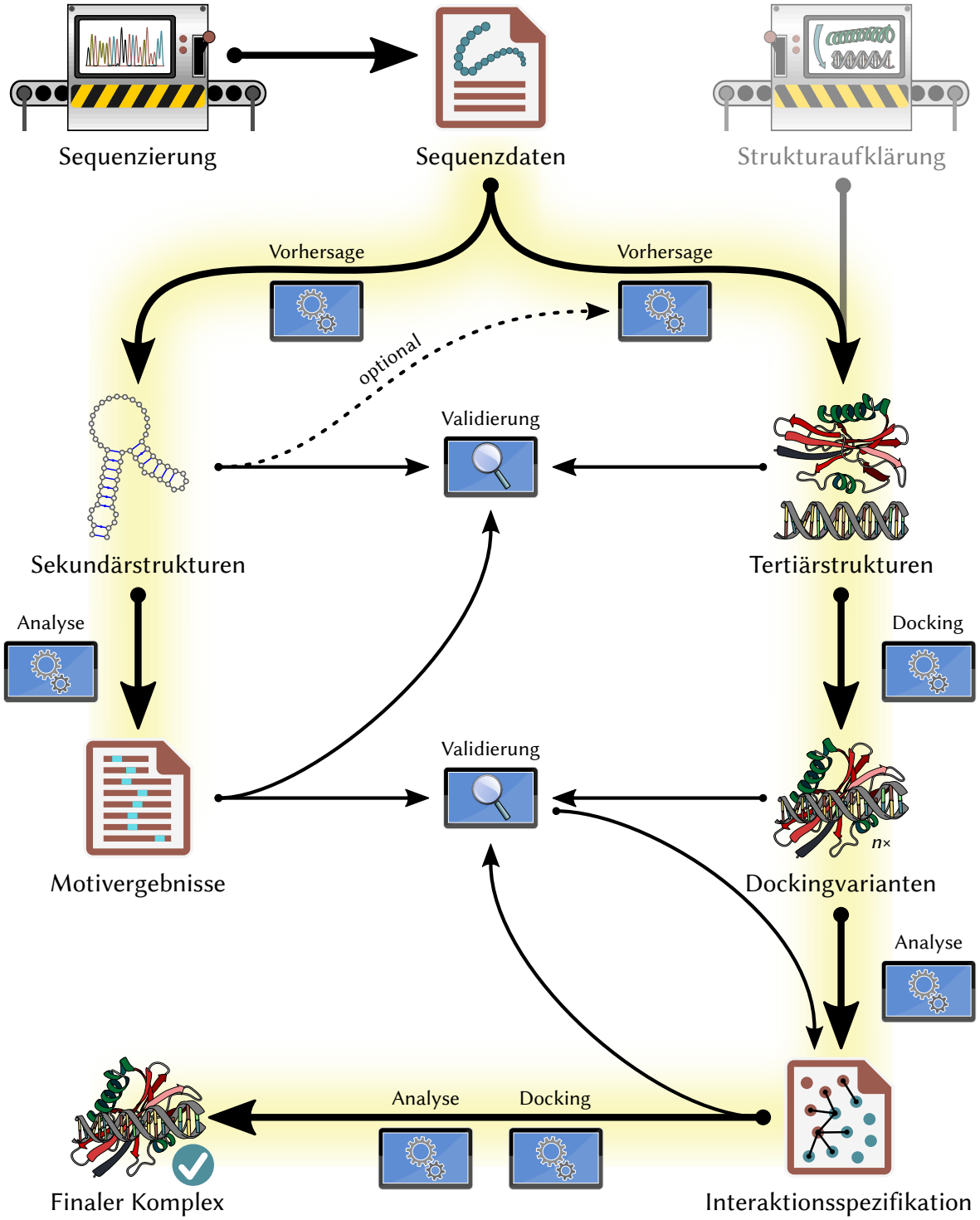


Abb. 6.7: Das Verfahrensprotokoll der bioinformatischen Analyse gibt ausgehend von den Sequenzinformationen der Bindungspartner zwei Hauptpfade (dicke Pfeile mit gelber Hinterlegung) vor. Aus einer hinreichenden Menge von Sequenzen werden im ersten Pfad (links) unter Zuhilfenahme vorhergesagter Sekundärstrukturen relevante Motive für die Bindung abgeleitet. Über einzelne, vorhergesagte Tertiärstrukturen kann im zweiten Pfad (rechts) mithilfe von Dockingsimulationen und elaborierten Bewertungsverfahren auf die Bindungsgeometrie des Komplexes aus Aptamer und Zielprotein geschlossen werden. Während des gesamten Prozesses ist der regelmäßige Abgleich der Teilergebnisse beider Pfade von hoher Wichtigkeit, da nur er die notwendige informationelle Rückkopplung gewährleistet, mit deren Hilfe die Schwächen der separaten Verfahren überwunden werden können. Der alternative Einstieg in den zweiten Pfad (rechts) über Strukturaufklärungsverfahren wurde in dieser Arbeit nicht realisiert.

Analysemethoden können sowohl mit den reinen Sequenzdaten als auch mit einer sekundärstrukturbasierten Modifikation selbiger ausgeführt werden, wobei die Gegenüberstellung der beiden wichtige Informationen bereitstellt.

Der zweite Hauptpfad verfolgt die Ableitung einer konkreten Bindungsgeometrie, anhand derer tatsächlich interagierende Residuen festgestellt werden können. Er bleibt jedoch aufgrund seiner Abhängigkeit von den Tertiärstrukturen auf vereinzelte Paare von Bindepartnern beschränkt. Das algorithmische Problem der Tertiärstrukturvorhersage ist durch die vielen Freiheitsgrade und komplexen Interaktionsmechanismen aus algorithmischer Sicht wesentlich schwerer als das der Sekundärstrukturvorhersage. So ist in den meisten Fällen bereits in der Vorhersage eine eigene Pipeline verschiedener Werkzeuge und Templatestrukturen notwendig, um zu einem verwertbaren Ergebnis zu gelangen. Ähnlich schwierig gestalten sich die Anforderungen einer experimentellen Strukturaufklärung, die in dieser Arbeit nur als Seitenvermerk Erwähnung findet. Bereits direkt nach Erhalt der Tertiärstrukturen geschieht ein erster Abgleich mit den Ergebnissen aus der Sekundärstrukturvorhersage. Sind die Strukturen hinreichend kompatibel, so wird die Lage der bereits ermittelten Sequenzmotive auf der Tertiärstruktur hinsichtlich ihrer Zugänglichkeit für den Bindepartner überprüft. Eine erste Dockingsimulation ohne Angabe von Bindungsspezifika liefert als Grundlage für die folgende Analyse eine große Anzahl sehr diverser Komplexvarianten. Mithilfe geeigneter Bewertungsschemata kann unter anderem auf energetischer Grundlage die Güte dieser Strukturen abgeschätzt werden. Basierend auf der damit einführten Ordnung erfolgt nun bei beiden Partnern die Ableitung von Präferenzen für mögliche Binderegionen. Dies geschieht stets mit Hinblick auf Gestalt und Lage der ermittelten Sequenzmotive, welche im Idealfall an den gefundenen Binderegionen beteiligt sind und gegebenenfalls deren Auswahl unterstützen. Die herausgestellten Regionen werden in eine formale Spezifikation präferierter Interaktionen überführt und somit dem Dockingsystem zugänglich gemacht. Auf der Basis dieser Spezifikation erfolgt eine weitere, diesmal gezielte Dockingsimulation, um den Strukturraum im Zielbereich höher aufzulösen. Mithilfe der bereits eingesetzten Bewertungs- und Analysemethoden wird anschließend der finale Komplex aus Aptamer und Zielmolekül ausgewählt und analysiert. Die gemeinsame Betrachtung aus seiner Bindungsgeometrie und der im ersten Hauptzweig festgestellten relevanten Sequenzmotive liefert schließlich wichtige Hinweise, die zur Optimierung der initialen SELEX-Bibliothek eingesetzt werden können.

6.3 Bioinformatische Analysen auf Basis der Primär- und Sekundärstruktur

Der erste Hauptpfad der bioinformatischen Analyse wird entsprechend dem entworfenen Verfahrensprotokoll in diesem Abschnitt behandelt. Als informationelle Grundlage dienten die Sequenzen und Sekundärstrukturen. In einer initialen Clusteranalyse wurde die Charakteristik der Anreicherung näher beleuchtet und das Vorhandensein von Subgruppen überprüft. Für die anschließende deskriptorbasierte Analyse wurde auf die methodischen Vorüberlegungen von Kapitel 3 zurückgegriffen. Hier stellten sich für die eingesetzte, regressionsbasierte Analysemethode n-Gramme als beste Beschreibungsform der Aptamersequenzen heraus. Schließlich wurde ein Mustersuchverfahren auf den Datensatz angewendet. Die Ergebnisse der unterschiedlichen Methoden wurden miteinander abgeglichen, um relevante Sequenzmuster für die Bindung des Zielproteins zu identifizieren und validieren.

6.3.1 Vorhersage der Sekundärstrukturen für die Aptamerkandidaten

Die in der Sekundärstruktur erfassten topologischen Restriktionen üben einen großen Einfluss auf die tatsächliche dreidimensionale Faltung eines Aptamers aus. Neben der essentiellen Bestimmung, welche Residuen für die Bindung des Zielmoleküls an der Oberfläche des Aptamers zugänglich sind, muss hier auch die innermolekulare Absättigung von Interaktionspotentialen beachtet werden. So finden sich Hinweise darauf, dass Bereiche ungebundener Nukleobasen in Aptameren eine höhere Tendenz zur Interaktion mit einem Zielprotein aufweisen können als bereits innermolekular gebundene Nukleobasen [488; 489]. Dies liegt an der Verfügbarkeit der zugehörigen Bindungsstellen für die intermolekulare Kopplung, da diese in *Loop*-Bereichen weder an Watson-Crick- noch an anderen, nicht-kanonischen Basenpaarungen teilnehmen.

Eingesetzte Software und Parametrisierung Unter den zahlreichen Softwareprodukten zur Vorhersage der Sekundärstruktur von Nukleinsäuren wurde aufgrund ihrer guten Reputation die Softwaresuite ViennaRNA [490] verwendet. Da die Grundeinstellung von ViennaRNA nur für die Vorhersage von RNA-Sekundärstrukturen parametrisiert ist, erforderte die Prozessierung von DNA-Sequenzen den Einsatz eines alternativen Energieparametersatzes. Die etablierten Energieparameter für DNA und RNA wurden in der Regel kooperativ bestimmt [491–493]. In dieser Arbeit kamen die Parameter von Mathews aus dem Jahr 2004 zum Einsatz. Im Gegensatz zu den bereits in Kapitel 3 untersuchten Promotoren ist im konkreten Anwendungsfeld der Aptamere jedoch ein deutlich höheres Maß an Variabilität der Umgebungsbedingungen zu erwarten. Da diese ebenfalls auf die Faltung der Aptamere Einfluss nehmen, birgt eine einfache Vorhersage der Sekundärstrukturen durch Energieminimierung ein erhöhtes Fehlerrisiko. In diesem Zusammenhang konnte gezeigt werden, dass die Energielandschaft im tatsächlichen Faltungsraum von Nukleinsäuren ähnlich wie bei Proteinen tendenziell trichterförmig beschaffen ist [494]. Tatsächlich finden sich auch in der Literatur Empfehlungen über den Einsatz energetisch suboptimaler Ensembles von Sekundärstrukturen [495; 496]. Dies ist besonders dann wichtig, wenn keine experimentellen Daten zur Verifikation oder Restriktion vorliegen [497] oder wenn ein Einfluss der Umgebungsbedingungen oder des Bindepartners auf die Faltung der Nukleinsäure nicht ausgeschlossen werden kann. So ist besonders bei Aptameren zu vermuten, dass sie aufgrund der Bindesituation in einer isoliert betrachtet energetisch suboptimalen Form vorliegen. Auch können Abweichungen in den approximierten Parametersätzen der Vorhersagemodelle bei Betrachtung eines suboptimalen Strukturensembles wesentlich besser toleriert werden [498]. Zur Vorhersage der Sekundärstrukturensembles wurde in dieser Arbeit die Software `RNAsubopt` der ViennaRNA Suite mit einer energetischen Toleranz von $2\,\mathrm{kcal\,mol^{-1}}$ verwendet.

Interpretation von Strukturensembles Die energetische Differenz zwischen gefaltetem und ungefaltetem Zustand eines Makromoleküls wird als seine freie Energie ΔG bezeichnet. Zwischen der freien Energie einer Sekundärstrukturausformung und der Häufigkeit ihres Auftretens in einem natürlichen Gemisch besteht ein definierter, stochastischer Zusammenhang, der in der Boltzmann-Verteilung beschrieben wird [498]. Damit kann für jede Struktur $s \in S$ des betrachteten Ensembles ausgehend von der freier Energie $\Delta G(s)$ entsprechend Formel 6.4 eine Wahrscheinlichkeit $p(s)$ für ihr Auftreten in einer natürlichen Verteilung abgeleitet werden. Sowohl die Temperatur T des Systems als auch die Boltzmann-Konstante k_B sind für die Berechnung notwendig, aber über alle Strukturen konstant. Über einen weiteren Parameter β, der im Normalfall mit dem Wert 1 belegt ist, kann die Verteilung zusätzlich gewichtet werden. Große Werte von β führen zu einer stärkeren Betonung der höher-energetischen Strukturen,

während kleine Werte den Einfluss der energetischen Bewertung zugunsten einer Gleichverteilung bei $\beta = 0$ schwächen. Die Zustandssumme Z berechnet sich über alle Strukturen in S und dient zur Normalisierung der Verteilung. Eine grundsätzliche Herausforderung bei der Arbeit mit Strukturensembles ist die Integration der zusätzlichen Strukturinformationen in die Bewertungsmodelle, die für einzelne Strukturen entwickelt wurden. Gemeinhin fließen die numerischen Ausprägungen abgeleiteter Kenngrößen in Form gewichteter Mittelwerte über dem Ensemble in die Berechnungen ein. Die während der Vorhersage berechneten freien Energien konnten genutzt werden, um Auftretenswahrscheinlichkeiten für die jeweiligen Strukturvarianten des Ensembles abzuleiten. Die Gewichtung erfolgte anschließend auf Basis dieser Werte.

$$p(s) = \frac{1}{Z} \cdot \mathrm{e}^{-\frac{\beta \cdot \Delta G(s)}{k_B \cdot T}} \qquad Z = \sum_{s \in S} \mathrm{e}^{-\frac{\beta \cdot \Delta G(s)}{k_B \cdot T}} \tag{6.4}$$

Einfluss von Primersequenzen Entsprechend des eingesetzten Sequenztemplates aus Tabelle 6.2 waren die randomisierten Insertsequenzen beidseitig mit konstanten Primersequenzen flankiert, die durch Basenpaarung mit der Insertsequenz auch während der Inkubationsphase des SELEX-Prozesses Einfluss auf die Tertiärstruktur der Aptamerkandidaten nehmen können. Für Aptamere kurzer [499–501] sowie mittlerer [502] Länge konnte der Einfluss dieser konstanten Primersequenzen auf die Ausformung der Struktur bereits experimentell bekräftigt werden. Auch wenn aus dieser Problematik Verfahren hervorgegangen sind, die während der Inkubationsphase mit verkürzten Primern oder gar primerfrei arbeiten [156; 158; 503], konnte auf Basis einer Simulation gezeigt werden, dass der tatsächliche Einfluss der konstanten Segmente eher gering bis moderat ausfällt. Eine mögliche Erklärung dafür ist, dass die Funktionsfähigkeit von Primersequenzen eingeschränkt wird, sobald diese zu komplex in die Struktur des Aptamers integriert werden. Das betroffene Aptamer ist dann im weiteren experimentellen Verlauf durch diese Benachteiligung nicht mehr konkurrenzfähig und scheidet aus [504]. Da diese Studie jedoch ausschließlich simulierte Strukturen analysierte und keinen markant negativen Effekt für die Betrachtung der Primersequenzen in der Sekundärstrukturvorhersage feststellen konnte, wurden die Primersequenzen in die Vorhersage der Sekundärstrukturen einbezogen.

Nutzungsvarianten der Sekundärstrukturinformation Die Sekundärstrukturinformation kann der weiteren algorithmischen Verarbeitung nun auf zwei unterschiedliche Arten zugänglich gemacht werden. Die erste Variante stellt eine starke Vereinfachung dar, welche für Teilbereiche der Sequenz die Wahrscheinlichkeiten erfasst, mit der sie vollständig auf *Loop*-Regionen gelegen sind. Dazu werden alle Strukturen des Ensembles selektiert, auf denen der betrachtete Sequenzbereich vollständig auf einer *Loop*-Region liegt. Der prozentuale Anteil dieser Strukturen an der Boltzmann-Verteilung ergibt die gesuchte Wahrscheinlichkeit, wie in Abbildung 6.8 am Beispiel einer kurzen Nukleotidsequenz gezeigt wird. Die so entstandene Annotation kann im weiteren als Auswahlkriterium oder Gewichtungsgrundlage verwendet werden. In der zweiten Variante wird aus jeder Sekundärstruktur des suboptimalen Ensembles jeweils basierend auf der eigentlichen Sequenz eine modifizierte Strukturseqenz erzeugt. Diese nutzt ein erweitertes Alphabet, das Großbuchstaben für ungepaarte und Kleinbuchstaben für gepaarte Nukleobasen einsetzt. Zwar geht auf diese Weise die genaue Zuordnung der Basenpaare verloren, die für die externe Bindung wichtige Information darüber, welche Nukleotide in der Bildung von Basenpaaren involviert sind, bleibt jedoch erhalten. Teilsequenzen können nun über das erweiterte Alphabet aus diesen Strukturseqenzen abgeleitet und entsprechend der Boltzmann-Verteilung gewichtet werden.

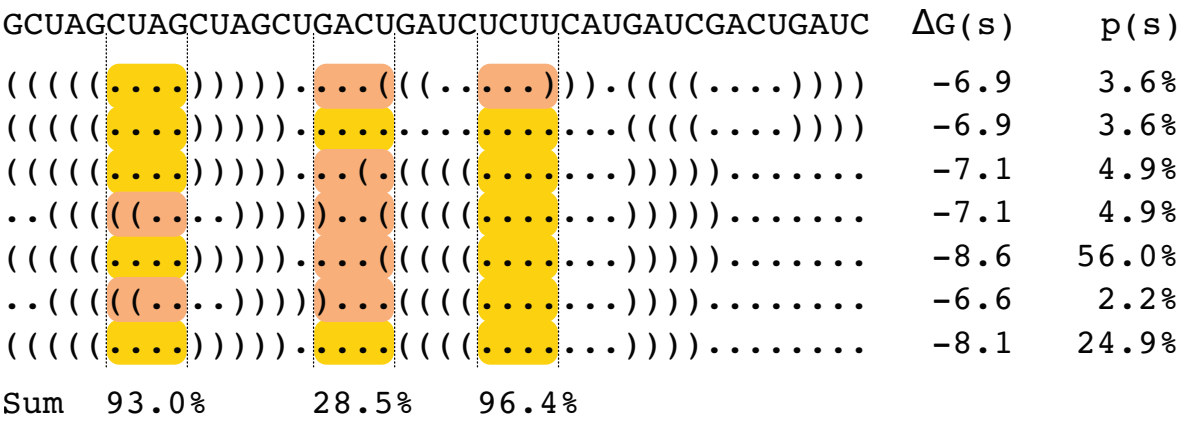


Abb. 6.8: Berechnung der Wahrscheinlichkeit für eine vollständige Positionierung spezifischer Teilsequenzen auf einer *Loop*-Region. Die obere Zeile zeigt die betrachtete Nukleotidsequenz. Die folgenden Zeilen enthalten jeweils eine energetisch suboptimale, vorhergesagte Sekundärstruktur in *Dot Bracket*-Notation. Punkte kennzeichnen dabei ungebundene Nukleobasen und zusammengehörende Klammerpaare repräsentieren Basenpaare. Jede dieser Strukturen s enthält eine Angabe der freier Energie $\Delta G(s)$ und der daraus errechneten Auftretenshäufigkeit $p(s)$. Eine gelbe Markierung hebt diejenigen Strukturvarianten hervor, auf denen die korrespondierende Teilsequenz vollständig auf einer *Loop*-Region liegt. Alle anderen Strukturvarianten der Teilsequenz sind orange hinterlegt.

6.3.2 Untersuchung der Anreicherung über die Clusteranalyse

Um die Anreicherung der Bibliothek über die Grenzen einer eindimensionalen Bewertungsgröße hinaus zu bewerten, wurde eine Clusteranalyse der unterschiedlichen Runden durchgeführt. Die Anreicherung manifestiert sich in einer solchen Analyse durch das Herausbilden und Wachsen einzelner Sequenzcluster. Für die Clusteranalyse war ein geeignetes Distanzmaß für die Aptamersequenzen notwendig. Häufig wird dieses aus den Ergebnissen eines paarweisen oder multiplen Alignments der Sequenzen abgeleitet. Durch die Fehleranfälligkeit der heuristischen Alignmentverfahren bestand jedoch die Gefahr der unbemerkten Verschleppung von Fehlern im weiteren Verfahren. Um dieser vorzubeugen, kam ein alignmentfreier Ansatz zur Anwendung.

Einfache Clusteranalyse Die einfach Clusteranalyse wurde auf Basis der n-Gramm-Zusammensetzung durchgeführt, da sich diese in Kapitel 3 als sehr gut geeignet für die Beschreibung von Nukleinsäuresequenzen herausstellte. Nachdem die absoluten Häufigkeiten $H_t(s)$ eines jeden n-Gramms t für jede Sequenz s bestimmt wurden, erfolgte die Festlegung der wiederholungsfreien Folge aller in den Sequenzen vorkommenden n-Gramme $T = (t_1, ..., t_k)$. Basierend auf dieser Folge wurde nun entsprechend Formel 6.5 für jede Sequenz ein vergleichbarer n-Gramm-Verteilungsvektor $\overrightarrow{G_s}$ abgeleitet. Anhand dieser Verteilungsvektoren wurde anschließend mithilfe standardisierter Distanzmaße wie dem Euklidischen Abstand $e(a,b)$ (Formel 6.6) die Distanzmatrix befüllt. Diese wurde dem *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA)-Algorithmus übergeben, entsprechend der Reihenfolge des UPGMA-Clusterbaumes angeordnet und zur visuellen Aufbereitung in Form einer Heatmap dargestellt.

$$\overrightarrow{G_s} = \begin{pmatrix} H_{t_1}(s) \\ \vdots \\ H_{t_k}(s) \end{pmatrix} \tag{6.5}$$

$$e(a,b) = \left\| \overrightarrow{G_a} - \overrightarrow{G_b} \right\| \tag{6.6}$$

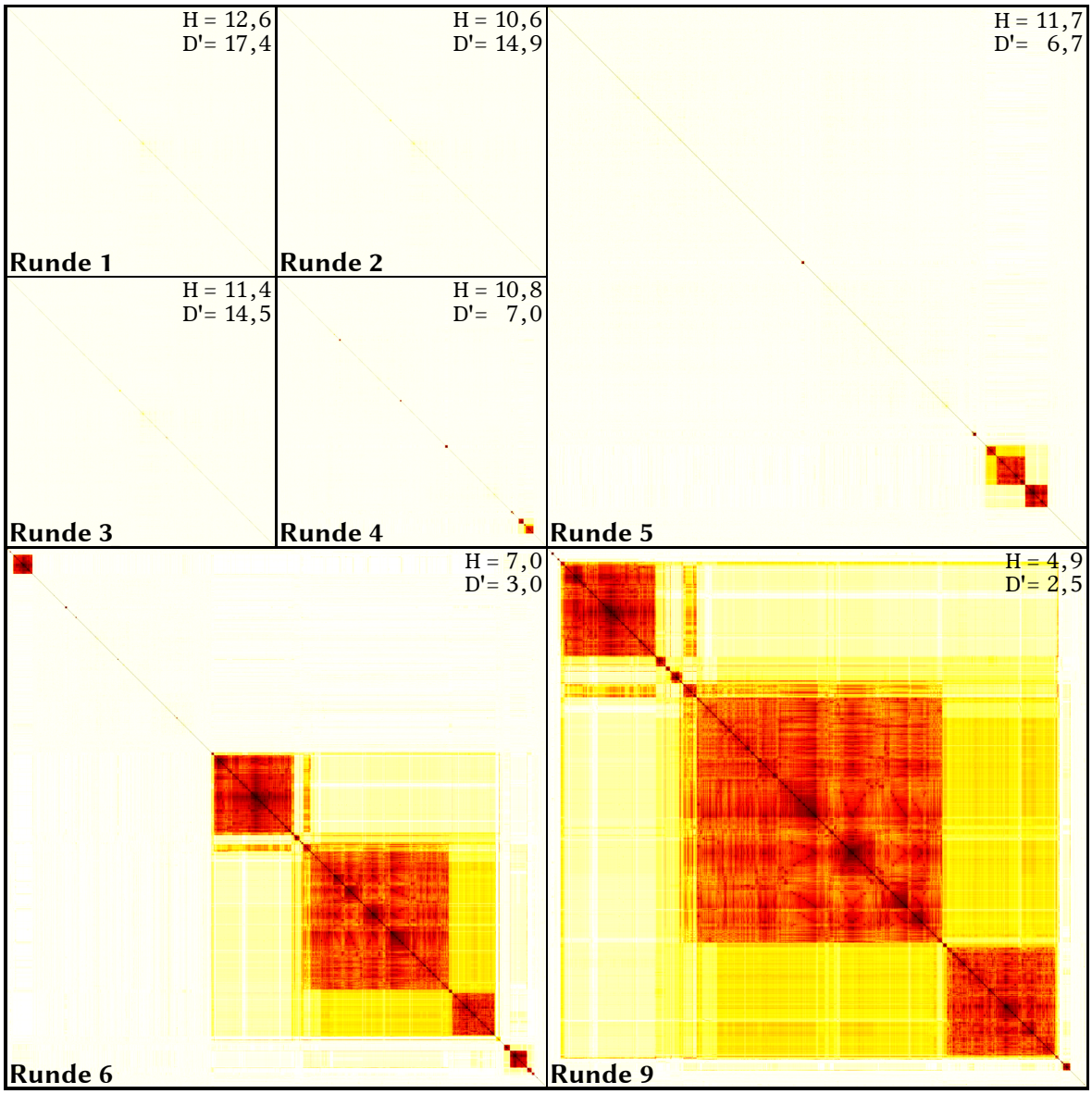


Abb. 6.9: Die jeweils 5000 häufigsten Sequenzen ausgewählter Runden wurden wie beschrieben einer Clusteranalyse unterzogen, als sortierte Heatmap visualisiert und mit den bereits bekannten Werten der Shannon-Entropie H und des modifizierten Simpson-Index D' annotiert. Die Einfärbung entspricht der Sequenzähnlichkeit im Bereich von weiß (keine Ähnlichkeit) über gelb und rot (leichte, bzw. hohe Ähnlichkeit) bis schwarz (identisch). Cluster sehr ähnlicher Sequenzen bilden sich in dieser Darstellungsform als rötliche Quadrate über der Hauptdiagonalen heraus. Beginnend mit Runde 4 ist eine Anreicherung in der Bibliothek zu beobachten, die sich bis Runde 8 entwickelt und anschließend stagniert.

Das beschriebene Verfahren wurde für eine ausgewählte Menge von Sequenzierungsrunden durchgeführt und in Abbildung 6.9 übersichtlich dargestellt. In den Runden 1 bis 3 sind dabei keine deutlich abgesetzten Cluster erkennbar, sondern lediglich eine geringe Anzahl sehr kleiner Bereiche mit schwacher sequenzieller Ähnlichkeit. Diese Beobachtung steht in Übereinstimmung mit der bereits erfolgten manuellen Kontrolle des Sequenzmaterials und bestätigt, dass dem Ansprechen der beiden numerischen Verfahren der Diversitätsbestimmung in diesem Bereich andere Ursachen zugrunde lagen. Beginnend mit Runde 4 sind erste kleine Sequenzcluster in der Abbildung zu erkennen, die übereinstimmend mit der Beobachtung der Diversität auf eine detektierbare Anreicherung hindeuten. Im Rahmen der Diversitätsbewertung wurde auf

eine starke Anreicherung einer kleinen Menge von Sequenzen geschlossen. Da im Clustering identische Sequenzen zusammengefasst wurden, standen hinter einzelnen Sequenzen dieser ersten Cluster höhere Kopienanzahlen, sodass die vorige Aussage auch hier bestätigt wurde. Bis hin zu Runde 9 wuchsen diese initialen Cluster in ihrer Größe stark an, sodass sie die Bibliothek schließlich dominierten. Der größte Schritt war hier zwischen den Runden 5 und 6 zu beobachten. Dies stimmt mit der Beobachtung der Diversität überein, die auf eine leichte Anreicherung einer großen Menge von Sequenzen hindeutete. Innerhalb der großen Cluster zeichnen sich in der Abbildung Bereiche höherer Ähnlichkeit ab, die eine Untergruppierung nahe legen. Neue Cluster bildeten sich neben diesen nur in sehr kleiner Zahl und geringer Ausprägung. Nach Runde 9 traten schließlich nur noch interne Veränderungen auf, die Charakteristik der bis dahin herausgebildeten Cluster blieb dann jedoch bis zur letzten Runde erhalten. Sowohl das verbliebene langsame Wachstum als auch die Stagnation wurden von beiden Diversitätsmaßen in gleicher Weise korrekt abgebildet.

Die Clusteranalyse lieferte im Vergleich zu den einfachen Diversitätsmaßen jedoch insgesamt einen höheren Detailgrad und damit höhere Verlässlichkeit in der Interpretation der Ergebnisse. Es empfiehlt sich daher, für die Auswertung auch auf die Clusteranalyse zurückzugreifen. Diese hat im vorliegenden Fall gezeigt, dass die Anreicherung qualitativ bereits mit Runde 8 abgeschlossen war. Ein Blick auf die vorhandenen Cluster konnte darüber hinaus bestätigen, dass die in den großen Clustern von Runde 12 vorhandenen Sequenzvertreter bereits in der ersten Ausbildung von Clustern in Runde 4 vorhanden waren, was ein Potential zur Verkürzung des experimentellen Ablaufs zeigt.

Clusteranalyse unter Einbezug der Sekundärstrukturen Da besonders *Loop*-Bereiche für die Interaktion zwischen Aptamer und Zielprotein wichtig sind [488; 489], wurde die Clusteranalyse mit einer Fokussierung auf derartige Bereiche wiederholt. Zur Herausstellung strukturell ungebundener Regionen wurden die vorhergesagten Sekundärstrukturen nach der ersten Nutzungsvariante eingebracht. Dies erforderte für jedes n-Gramm die Berechnung der Wahrscheinlichkeit für seine vollständige Positionierung auf einer *Loop*-Region. Die Verteilungsvektoren folgen dem gleichen Aufbau wie bereits beim einfachen Clustering, basierten jedoch nicht mehr auf den Auftretenshäufigkeiten, sondern auf den summierten Wahrscheinlichkeiten. Abbildung 6.10 zeigt das Ergebnis dieses erweiterten Clustering-Laufes im Vergleich zum unveränderten Verfahren für die letzte sequenzierte Runde. Während die Cluster im einfachen Verfahren optisch sehr deutlich und scharf trennbar in der Abbildung hervortreten, so ergibt sich durch das erweiterte Verfahren ein eher verwaschenes Bild. Die lagebasierte Gewichtung der n-Gramme wirkte gleichzeitig als starker Einschnitt in die Grunddatenmenge der Verteilungsvektoren, der sich massiv auf die Ähnlichkeitsbeiträge der einzelnen n-Gramme niederschlug. So verloren gleiche Sequenzabschnitte bei gemeinsamer Lage auf einem doppelsträngigen Bereich ihre Ähnlichkeit, während ihre unterschiedliche Positionierung sogar zu einer Zunahme der Unähnlichkeit führte. Im gleichen Zuge erreichte die Gewichtung jedoch auch die Hervorhebung all jener Ähnlichkeiten, die für die Bindung des Aptamers mit dem Zielprotein von besonderer Wichtigkeit sind. Die häufigste in der letzten Runde vorkommende Sequenz (siehe Tabelle 6.6) wurde nach Durchlauf beider Varianten des Clusteringverfahrens jeweils einem unterteilten Cluster zugeordnet. In Abbildung 6.10 wurden die entsprechenden Cluster mit einem sehr ähnlichen Kern- und einem ausreichend ähnlichen Erweiterungsbereich durch eine blaue Markierung am jeweils äußeren Bildrand gekennzeichnet. Der Kernbereich beschränkt sich dabei auf Sequenzen sehr hoher gegenseitiger Übereinstimmung (<4 einzelne Mutationen),

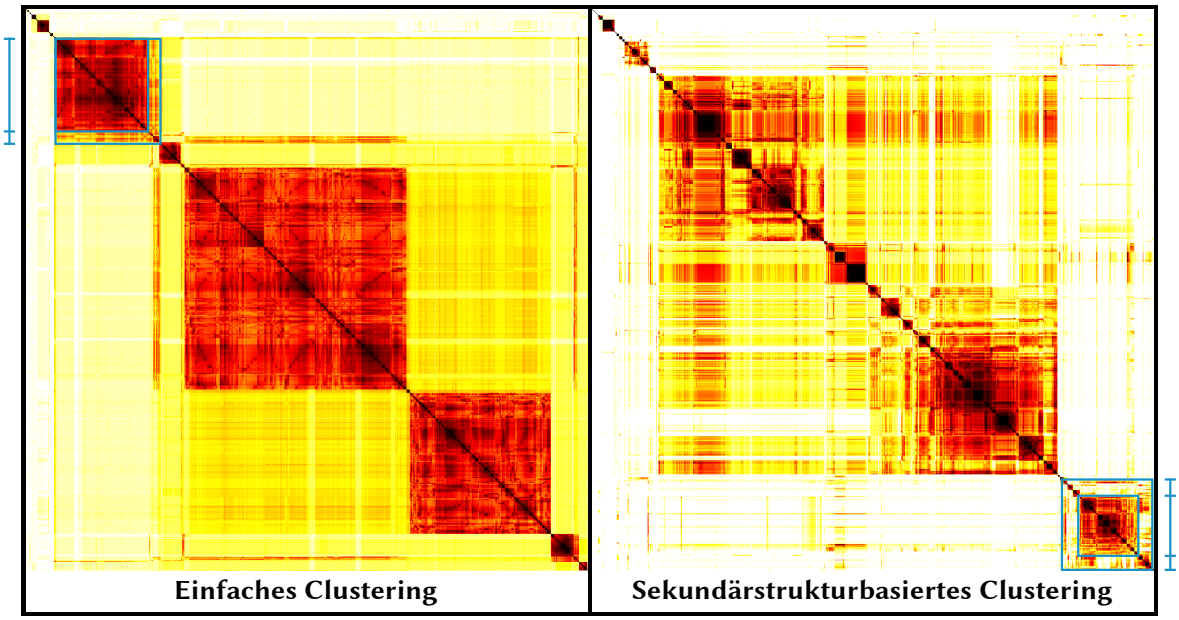


Abb. 6.10: Die jeweils 1000 häufigsten Sequenzen der Bibliothek nach Runde 12 wurden sowohl der einfachen (links) als auch der sekundärstrukturbasierten (rechts) Clusteranalyse unterzogen. Das Farbschema wurde dabei von Abbildung 6.9 übernommen. Die geschachtelten, blauen Markierungen an den äußeren Abbildungsrändern korrespondieren mit den gleichfarbigen Rahmen in der Heatmap. Sie kennzeichnen jeweils das System aus Kerncluster und Erweiterungsbereich, in dem sich die häufigste Sequenz der letzten Runde befindet, und gehören damit verfahrensübergreifend zusammen. In der Darstellung ist durch Hinzunahme der Sekundärstrukturinformationen die grundlegende Charakteristik der Anreicherung, wenn auch weniger klar, erhalten geblieben.

die im einfachen Clustering etwa 25 % und beim erweiterten noch rund 21 % des gesamten Sequenzpools ausmachten. Im erweiterten Cluster kommen zu diesen noch weitere Sequenzen mit teilweisen Übereinstimmungen in der ersten Sequenzhälfte hinzu.

In Abbildung 6.10 zeigt sich, dass die Gesamtstruktur des Clusterings trotz der Verluste in Homogenität und Trennschärfe sowie wenigen Verschiebungen nicht verloren gegangen ist. Auch unter Anwendung des strengeren, auf der Sekundärstruktur basierenden Ähnlichkeitskriteriums konnte eine Anreicherung in der Bibliothek festgestellt werden, die mit der des einfachen Clusterings korrespondiert. Basierend auf diesen Beobachtungen kann angenommen werden, dass sich in der Form der gefundenen Cluster tatsächlich relevante Ähnlichkeiten in *Loop*-Regionen in der experimentellen Bibliothek angereichert haben.

6.3.3 Untersuchung der n-Gramm-Zusammensetzung

Da sich n-Gramme als beste Beschreibungsform der Nukleinsäuresequenzen herausgestellt hatten, wurde im Zuge der bioinformatischen Analyse nun die Zusammensetzung der Sequenzdaten aus n-Grammen in Hinblick auf mögliche Beiträge zur Affinität der Aptamerkandidaten untersucht. Neben der Hauptanalyse der finalen Runde erfolgte eine einfache Betrachtung mit Hinblick auf die Bibliotheksentwicklung. Die Übernahme des Vorgehens aus Kapitel 3 umfasste die PLS-Regression mitsamt *Feature Selection*, erforderte aber die Sicherstellung zweier Anforderungen. Als Zielgröße für die Optimierung war eine Affinitätsinformationen zum Zielprotein notwendig, die aus der Sequenzierung nicht hervorging. Da das einzelne Synthetisieren aller infrage kommenden Aptamerkandidaten mit anschließender experimenteller Bestimmung der Affinität keine Alternative für eine zügige Analyse darstellte, musste die Affinität aus der Häufigkeitsverteilung der Sequenzdaten geschätzt werden. Die zweite Anforderung betraf die Größe

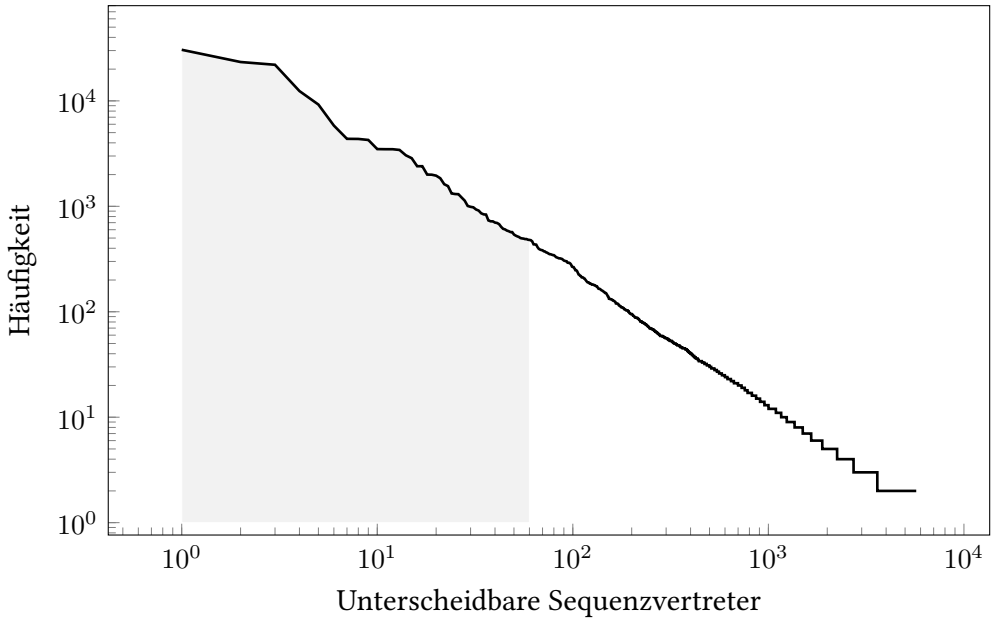


Abb. 6.11: Die sortierte Häufigkeitsverteilung aller unterscheidbaren Sequenzvertreter der letzten sequenzierten Runde wurde logarithmisch aufgetragen. Neben wenigen sehr häufigen Vertretern fällt die Verteilung exponentiell ab und läuft schließlich in einer großen Menge sehr schwach vertretener Sequenzen aus. Die Auswahl für die weitere Analyse durch die n-Gramm-Methode ist hellgrau hinterlegt. Sie umfasst mit den häufigsten 60 Sequenzvertretern circa 72 % des gesamten Datensatzes.

des zu analysierenden Datensatzes. Eine Verminderung war jedoch nicht nur aus Gründen der Laufzeit notwendig, sondern hauptsächlich infolge der negativen Effekte des informationellen Rauschens, das durch die zahlreichen, sehr niedrig angereicherten Sequenzen in die Daten kam. So wirken sich Sequenzierungsfehler und unvermeidliche statistische Einflüsse bei der Auswahl der zu sequenzierenden Probe aus der Gesamtbibliothek bei niedriger Anreicherung besonders stark auf die berechneten Affinitäten aus. Bedingt durch die experimentelle Anordnung können ferner kleine Mengen an Oligonukleotiden in der Bibliothek verbleiben, die trotz Negativselektionsrunden nicht mit dem Zielprotein, sondern mit den genutzten Hintergrundmaterialien interagieren. Bei der Auswahl einer zu analysierenden Teilbibliothek wurden diese Umstände beachtet und niedrig angereicherte Sequenzen verworfen.

Aufbereitung der finalen Bibliothek Besondere Beachtung kam der letzten sequenzierten Bibliothek zu, da sich der Großteil der Analysen auf diese bezog. Der zugehörige Datensatz trug die Bezeichnung `FH-6-12` und zählte ursprünglich 348 892 Sequenzen (siehe Tabelle 6.4). Nach Filterung und Aggregation ergab sich hieraus ein Datensatz von 5703 unterscheidbaren Sequenzen mit Auftretenshäufigkeiten von 2 bis 30 516. Die in Abbildung 6.11 dargestellte Verteilung dieser Vorkommenshäufigkeiten zeigt, dass ein großer Teil der Sequenzen in sehr geringer Zahl vorkam. Entsprechend der Vorüberlegungen wurde das Rauschen durch Verwerfen niedrig angereicherter Sequenzen vermindert. Die Auswahl der häufigsten 60 Sequenzen deckte mit circa 72 % bereits den Großteil des Datensatzes ab. Um Verzerrungen der Ergebnisse zu verhindern, wurden für die Durchführung der Analysen ebenfalls 60 Negativsequenzen mit gleicher Nukleobasenkomposition durch Permutation der Positivsequenzen erzeugt. Durch die gleiche Komposition ist die Beschreibung durch 1-Gramme gegenstandslos geworden.

Die Sekundärstrukturinformationen wurden aufgrund ihrer bereits festgestellten Relevanz auch in die Untersuchung der n-Gramm-Beteiligungen eingebracht. Die Art der Einbringung wich jedoch von der Vorgabe in Kapitel 3 ab, da bei Aptameren im Gegensatz zu Promotoren deutlich größere Schwankungen der Umgebungsbedingungen vorherrschen. Diese beeinflussen die energetischen Zustände der Aptamere, sodass die Struktur mit minimaler freier Energie unter optimalen Bedingungen nicht zwangsläufig auch in der komplexen Umgebung vorherrscht. Zur Ableitung der n-Gramme wurden daher suboptimale Sekundärstrukturensembles der Aptamere entsprechend der zweiten Nutzungsvariante verwendet. Um die Informationen aller vorhergesagten suboptimalen Strukturensequenzen zu nutzen, waren einige Anpassungen am Verarbeitungsmechanismus erforderlich. So führte das temporäre Anfügen der Primersequenzen während der Vorhersage nach deren Entfernung zu strukturellen Mehrfachvorkommen im *Insert*-Bereich der Sequenzen. Ein Schritt zur Zusammenfassung dieser Multiplizität wurde eingerichtet.

Ableitung der Affinität Initial wurde überprüft, ob sich die relativen Häufigkeiten der Sequenzen direkt als Schätzung ihrer Affinitäten zum eingesetzten Zielprotein eigneten. Bei der probeweisen Durchführung des Analyseverfahrens zeigte sich jedoch, dass die Hinzunahme sinnvoller Deskriptoren kaum Effekt in der Optimierung zeigte und der Permutationstest auf die Unzuverlässigkeit der Ergebnisse hinwies. Zur Ableitung einer Affinitätsinformation wurde daher der Weg über die Boltzmann-Verteilung beschritten, die einen Zusammenhang zwischen Affinität und Auftretenshäufigkeiten eines Aptamers in der natürlichen Mischung der experimentellen Bibliothek beschreibt. Über die Invertierung der aus Formel 6.4 bereits bekannten Relation konnte daher aus der Häufigkeitsverteilung der Bibliothek auf die zugrundeliegenden Affinitäten geschlossen werden. Obwohl die unbekannte kanonische Zustandssumme Z die vollständige Invertierung der Relation verhinderte, erlaubte ihre konstante Natur die Ableitung von Differenzenergiebeträgen $\Delta E(s)$ nach Formel 6.7. Aus den oben genannten Gründen beschränkte sich die Betrachtung auf die Häufigkeitsverteilung $h^k(S^k)$ der häufigsten k Sequenzen $S^k = (s_1, \ldots, s_k)$ der betrachteten Bibliothek. Als Maß für die Affinität konnte jedoch nur eine absolute Größe der Bindungsenergie dienen, sodass ein geeigneter Referenzzustand gefunden werden musste. Hierfür wurde die am seltensten vorkommende Sequenz $\arg\min_{s \in S} h(s)$ bezogen auf die Häufigkeitsverteilung $h(S)$ aller Sequenzen des Datensatzes verwendet. Durch Subtraktion ergab sich die absolute, approximierte Bindungsenergie $E(s)$, deren Berechnung nach Formel 6.8 erfolgte. In Tabelle 6.7 findet sich eine Übersicht mit den gewählten 60 Sequenzen der finalen Bibliothek und den entsprechend abgeleiteten Affinitätswerten. Die additive Natur dieser n-Gramm-weisen, energetischen Beiträge kommt dem natürlichen Prinzip der molekularen Wechselwirkungen sehr nahe, da auch diese sich aus Einzelbeträgen zusammensetzen.

$$\Delta E(s) = -\frac{k_B T}{\beta} \cdot \ln h^k(s) \tag{6.7}$$

$$\begin{aligned} E(s) &= -\frac{k_B T}{\beta} \left(\ln h^k(s) - \ln \left(\min_{\sigma \in S} h(\sigma) \right) \right) \\ &= -\frac{k_B T}{\beta} \cdot \ln \left(\frac{h^k(s)}{\min\limits_{\sigma \in S} h(\sigma)} \right) \end{aligned} \tag{6.8}$$

Da für die Bewertung der Affinität von Negativsequenzen aus mangelnder Datenlage kein realer Wert bestimmt werden konnte, wurde ungeachtet der sequenziellen Unterschiede für alle Negativsequenzen ein konstanter Wert festgelegt. Ist eine Sequenz aufgrund ihrer Bindungsun-

Tab. 6.7: Tabellarische Darstellung des Sequenzdatensatzes für die n-Gramm-Analyse. Der Datensatz setzt sich aus den 60 häufigsten Sequenzen der letzten experimentellen Runde zusammen, die mit circa 72 % bereits den Großteil der sequenzierten Bibliothek abdecken. Die Sequenzen sind entsprechend der jeweiligen Nukleobasen eingefärbt. Bei dem angegebenen Affinitätswert (Aff.) handelt es sich um eine einheitenlose Schätzung auf Basis der Sequenzhäufigkeiten. Die Sequenzen sind von der höchsten zur niedrigsten Affinität sortiert.

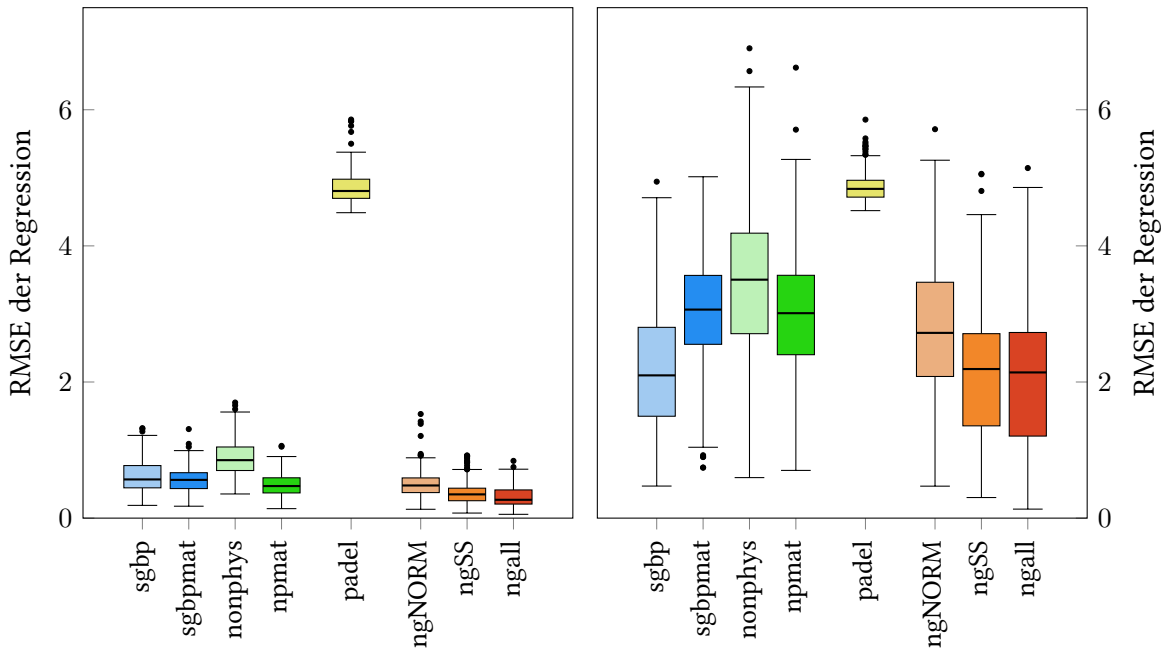
Nr	Sequenz	Aff.
1	GTCTGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGGC	6.1409
2	CATGGAGGTGGTCCGGGTCCCATGGTCAGTCACAAACTAGTCACCAAGG	5.9764
3	CATGGAGGTGGTCCGGGTCCCATGGTCAGCCACAAACTAGTCACCAAGG	5.9394
4	CATTCGCGGAACTATGTGGAGGTGGTCCGGGACCCGCAAAGTTTCGGGC	5.5858
5	GTCTGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGAC	5.4043
6	AGTGAGGAGAGGTAGGCTTACTGGGAGGTGGTCCGGGGCCCCAGTACAG	5.1213
7	CATTCGCGGAACTATGTGGAGGTGGTCCAGGACCCGCAAAGTTTCGGGC	4.9433
8	CATGGAGGTGGTCCGGGTCCCATGGTCAGTCACAAACTAGTCACCGAGG	4.9417
9	CATGGAGGTGGTCCGGGTCCCATGGTTAGTCACAAACTAGTCACCAAGG	4.9280
10	CATGGAGGTGGTCCGGGTCCCATGGTCAGTCACAAACTAGTCACCAGGG	4.8052
11	ACGCATGGTTTGCCACTACCTGGACACATGGTCGGGCCGTCTGAGGGTA	4.8036
12	GTCTGTAGCAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGGC	4.8034
13	GCGTAGTCCCAGCACGGTCTGCCAAAAAGCATGGAGGTGGTCCGGGTCC	4.7938
14	CATGGAGGTGGTCCGGGTCCCATGGTCAGTCACAAACTAGTCATCAAGG	4.7218
15	CATGGAGGTGGTCCGGGTCCCATGGTCAGCCACAAACTAGTCACCAGGG	4.6827
16	ACGTAGTCCCAGCACGGTCTGCCAAAAAGCATGGAGGTGGTCCGGGTCC	4.5747
17	GTCTGTAGCAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGAC	4.5747
18	CATGGAGGTGGTCCGGGTCCCATGGTTAGCCACAAACTAGTCACCAAGG	4.4624
19	CATGGAGGTGGTCCGGGTCCCATGGTCAGTCACAAACTAGTTACCAAGG	4.4609
20	CATGGAGGTGGTCCGGGTCCCATGGTCAGCCACAAACTAGTCATCAAGG	4.4481
21	CATGGAGGTGGTCCGGGTCCCATGGTCAGCCACAAACTAGTCACCGAGG	4.4131
22	GTTTGTAGCAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGGC	4.3330
23	GTCCGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGGC	4.3074
24	GTCTGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTGTCAGGC	4.2084
25	CATGGAGGTGGTCCGGGTCCCATGGTTGGCCACAAACTAGTCACCAAGG	4.2009
26	ATCCCCTCTGGTATCGCGTAAGACGACGTGCTGGCCATCGTGTGCTCTA	4.1985
27	CATGGAGGTGGTCCGGGTCCCATGGTCAGTCACAAACTAGTCGCCAAGG	4.1540
28	CATGGAGGTGGTCCGGGTCCCATGGTCAGCCACAAACTAGTTACCAAGG	4.1119
29	CATGGAGGTGGTCCGGGTCCCATGGTCAGCCGCAAACTAGTCACCAAGG	4.0404
30	TATGGAGGTGGTCCGGGTCCCATGGTTAGTCACAAACTAGTCACCAAGG	4.0287
31	GTCTGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATTAGGC	4.0218
32	CATGGAGGTGGTCCGGGTCCCATGGTCAGTCACAAACCAGTCACCAAGG	3.9928
33	CATGGAGGTGGTCCGGGTCCCATGGTCGGCCACAAACTAGTCACCAAGG	3.9781
34	GTCTGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGGT	3.9419
35	ATGCATGGTTTGCCACTACCTGGACACATGGTCGGGCCGTCTGAGGGTA	3.9274
36	GTCTGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATTAGAC	3.9245
37	GCGTAGTCCCAGCACGGTCTGTCAAAAAGCATGGAGGTGGTCCGGGTCC	3.8459
38	ACGTAGTCCCAGCATGGTCTGCCAAAAAGCATGGAGGTGGTCCGGGTCC	3.8341
39	AGTGAGGAGAGGTAGGCTTACTGGGAGGTGGTCCGGGGCCCCAGTACGG	3.8332
40	CATGGAGGTGGTCCGGGTCCCATGGTCAGCCACAAACTAGTCGCCAAGG	3.8167
41	GCGTAGTCCCAGCATGGTCTGCCAAAAAGCATGGAGGTGGTCCGGGTCC	3.8114
42	CATGGAGGTGGTCCGGGTCCCATGGTTAGTCACAAACTAGTCACCGAGG	3.7953
43	CATGGAGGTGGTCCGGGTCCCATGCTTATTCTTGTCTCCCAGATCGGAA	3.7597
44	GTGTAGTCCCAGCACGGTCTGCCAAAAAGCATGGAGGTGGTCCGGGTCC	3.7331
45	GCCTGTAGTAGGGAGGATGGTCCGGGGCCCCGAGACGACGTTATCAGGC	3.7250
46	CATGGAGGTGGTCCGGGTCCCATGGTCAGCCACAAACCAGTCACCAAGG	3.7095
47	CATGGAGGTGGTCCGGGTCCCATGGTTAGCCACAAACTAGTCATCAAGG	3.7011
48	CATGGAGGTGGTCCGGGTCCCATGGTTAGTCACAAACTAGTCACCAGGG	3.6893
49	ACGTAGTCCCAGCACGGTCTGTCAAAAAGCATGGAGGTGGTCCGGGTCC	3.6882
50	TATGGAGGTGGTCCGGGTCCCATGGTTAGCCACAAACTAGTCACCAAGG	3.6560
51	GCGTAGTCCCAGCACGGTCTGCCAAGAAGCATGGAGGTGGTCCGGGTCC	3.6422
52	GCGTAGTCCTAGCACGGTCTGCCAAAAAGCATGGAGGTGGTCCGGGTCC	3.6316
53	CATGGAGGTGGTCCGGGTCCCATGGTTAGTCACAAACTAGTCATCAAGG	3.6232
54	CATGGAGGTGGTCCGGGTCCCATGGTCAGTCACAAACTAGTCACCGGGG	3.6086
55	GCGTAGTCTCAGCACGGTCTGCCAAAAAGCATGGAGGTGGTCCGGGTCC	3.6037
56	GCGTAGTCCCAGTACGGTCTGCCAAAAAGCATGGAGGTGGTCCGGGTCC	3.6012
57	ACTAGTCCGCGTGTTATGGTCCCATAACCAACTGTTCGCGGTTTTATAG	3.5949
58	TGAGCTGTAGGAGGTGGTCCGGGACCCTGGTGCGAGGGGGGAAGTGAAC	3.5949
59	CATGGAGGTGGTCCGGGTCCCATGGTTAGCCACAAACTAGTCACCAGGG	3.5860
60	CATGGAGGTGGTCCGGGTCCCATGGTTAGCCACAAACTAGTCACCGAGG	3.5835

fähigkeit in der Bibliothek nicht vorhanden, so ergibt sich durch die Anwendung von Formel 6.8 der Grenzwert $\lim_{h^k(s)\to 0} E(s) = -\infty$ als energetische Bewertung. Für den praktischen Einsatz war dieser Grenzwert ungeeignet, sodass der konstante Affinitätswert für negative Sequenzen auf -5 festgelegt wurde. Der Wert gibt mit seiner großen Differenz zu den tatsächlichen Affinitäten ein Gegengewicht für das Regressionsverfahren, ohne eine zu große Verzerrung hervorzurufen. Auf die explizite Nennung der 60 zufällig permutierten Negativsequenzen wird an dieser Stelle verzichtet.

Verifikation des Beschreibungsverhaltens der n-Gramme Die Untersuchung der Deskriptoren in Kapitel 3 wurde repräsentativ an einem Datensatz von Promotorsequenzen durchgeführt. Aufgrund des gemeinsamen Wirkungsprinzips funktioneller Nukleinsäuren über die molekulare Erkennung wurde begründeterweise von einer Verallgemeinerbarkeit der Erkenntnisse ausgegangen. Die Korrektheit dieser Annahme wurde für Aptamere explizit überprüft. Um die gleichen Ausgangsbedingungen herzustellen, wurde dazu eine Auswahl der häufigsten Aptamersequenzen von der Größe des in Kapitel 3 verwendeten Promotorendatensatzes getroffen und durch entsprechend randomisierte Negativsequenzen ergänzt. Für diesen Analogdatensatz wurde die in Kapitel 3 bereitgestellte Pipeline aus Vorverarbeitung, numerischer Beschreibung und Regressionsverfahren vollständig ausgeführt, was neben den n-Grammen auch die anderen Deskriptoren umfasste. Die vorgeschlagenen Modifikationen in der Handhabung der Sekundärstrukturinformation wurden jedoch nicht eingepflegt, um die Vergleichbarkeit der Ergebnisse zu wahren.

Das Ergebnis nach der *Feature Selection* wird in einer Kurzform in Abbildung 6.12 dargestellt und umfasst auch die Negativprobe durch Permutation der Zielgröße. Bei Verwendung des Aptamerdatensatzes zeigten sich bei allen Deskriptorensets außer bei dem Set PADEL, welches keine positionelle Information enthält, sehr geringe Modellfehler. PADEL außen vor gelassen, waren die Unterschiede zwischen den verbliebenen Sets zwar eher gering, es zeichnete sich aber dennoch ein merklicher Vorsprung in der Genauigkeit der n-Gramm-basierten Beschreibung ab. Dies galt besonders bei der Kombination von n-Gramm-Deskriptoren unterschiedlicher Länge (nicht gesondert abgebildet, siehe dann Abbildung 6.13) als auch beim Einsatz der Sekundärstrukturinformationen. Diese Beobachtung bestätigte sowohl die Wichtigkeit der positionellen Information in der Beschreibung der Aptamere als auch die hohe Eignung von Deskriptoren auf n-Gramm-Basis als Abbildung natürlich vorkommender Bindemotive. Im Rahmen eines Permutationstests konnte gezeigt werden, dass die angewendete Optimierungsstrategie auf einem Datensatz mit falsch zugeordneten Affinitätswerten ausschließlich Modelle generiert, deren Modellfehler um mehr als eine Größenordnung über der Referenz liegen. Es konnte damit sichergestellt werden, dass es sich bei den vorliegenden Regressionsmodellen nicht um Ergebnisse einer Überanpassung handelte. Die hervorragende Beschreibungsfähigkeit der n-Gramme ließ sich daher auch auf Aptamersequenzen übertragen.

Analyse der finalen Bibliothek Das in Kapitel 3 beschriebene Analyseverfahren wurde für die n-Gramm-Deskriptoren auf die aufgearbeitete finale Bibliothek angewendet. Abbildung 6.13 gibt eine visuelle Übersicht der erreichten Ergebnisse von Optimierung und Permutationstest. Die Beschreibung mit 1-Grammen zeigte das typische Verhalten einer Zufallsgröße ohne kontextuellen Bezug, da diese bei gleicher Nukleobasenkomposition gegenstandslos war. Auch die Hinzunahme der Sekundärstruktur in Form des erweiterten Alphabetes führte zu keinem signifikanten Informationsgewinn (nicht abgebildet). Mit steigender Länge der n-Gramme zeigen die

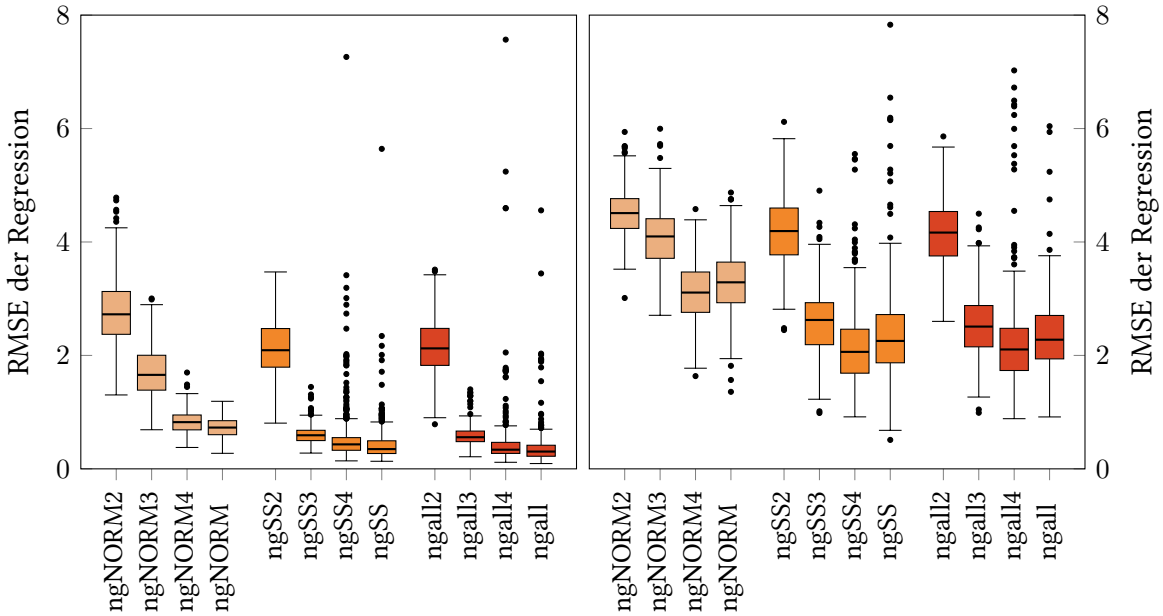


Originaldatensatz **(a)** **(b)** Permutation der Zielgröße

Abb. 6.12: Überblick über eine repräsentative Auswahl der Aptamer-Beschreibungssets (siehe Kapitel 3) nach erfolgter *Feature Selection*. Vergleichsgrößen sind die kreuzvalidierten RMSE-Werte der PLS-Regressionsmodelle, die auf den jeweiligen Beschreibungssets trainiert wurden. Im Originaldatensatz (a) zeigt sich das erwartet gute Verhalten der Beschreibung durch n-Gramme. Nach zufälliger Permutation der Zielgröße (b) zeigen sich deutlich größere Modellfehler.

abfallenden mittleren Modellfehler in allen drei Gruppen der Sekundärstrukturnutzung (`NORM`, `ss`, `all`) in der Abbildung die wachsende Aussagekraft der einhergehenden Sequenzfragmente an. Bereits bei einer Länge von 4 stellte sich jedoch die Sättigung dieser Verbesserung ein, da die zunehmende Spezifität und abnehmende Generalisierungsfähigkeit längerer n-Gramme kaum mehr einen informationellen Gewinn für das Verfahren darstellte. Im Vergleich bei fester Länge zeigen sich in der Abbildung Verbesserungen der Modellgüte durch die Hinzunahme der Sekundärstrukturinformationen, die jedoch im Sättigungsbereich abnehmen. Die große Anzahl von Deskriptoren, die sich aus der Kombination von langen Fragmenten und Sekundärstrukturinformationen ergab, beeinträchtigte das Optimierungsverfahren, sodass es in diesen Fällen vermehrt zu Ausreißern kam. Basierend auf der Zielstellung einer generalisierungsfähigen, homogeneren Vorhersagequalität eignete sich für die weitere Untersuchung daher das gemischte Deskriptorenset `ngall3` am besten. Zwar zeigt sich auch in der permutierten Gegenprobe eine Abnahme der Modellfehler nach dem eben beschriebenen Muster, jedoch spielte sich diese in einer deutlich höheren Größenordnung >2 ab. Die Gegenprobe wies ferner einen größeren Anteil stärkerer Ausreißer auf, die den gewählten Darstellungsbereich zum Teil übersteigen. Eine Überanpassung durch das genetische Optimierungsverfahren der *Feature Selection* hat daher nicht stattgefunden.

Für eine Auswertung der konkreten n-Gramm-Beteiligungen wurde das Regressions- und Selektionsverfahren für das gewählte Deskriptorenset `ngall3` unter strengeren Optimierungsbedingungen erneut durchgeführt. Zu diesem Zweck wurde nicht nur die Populationsgröße erhöht, sondern auch das Abbruchkriterium derart angepasst, dass längere Phasen mit sehr geringer oder keiner Verbesserung toleriert werden. Von den insgesamt 93 durch die *Feature Se-*



Originaldatensatz **(a)** **(b)** Permutation der Zielgröße

Abb. 6.13: Überblick über die erreichbaren Modellfehler bei Beschreibung der Aptamere durch n-Gramm-Deskriptoren nach erfolgter *Feature Selection*. Vergleichsgrößen sind die kreuzvalidierten RMSE-Werte der PLS-Regressionsmodelle, die auf den jeweiligen Beschreibungssets trainiert wurden. Die Einfärbung unterscheidet die drei Gruppen mit unterschiedlichem Nutzungsgrad der Sekundärstrukturen. Im Originaldatensatz (a) sinken die Modellfehler sowohl mit der Länge der n-Gramme als auch mit der Nutzung der Sekundärstrukturinformationen. Nach zufälliger Permutation der Zielgröße (b) zeigen sich deutlich größere Modellfehler.

lection gewählten Deskriptoren entfiel mit $85\,\%$ der größte Teil auf sekundärstrukturabhängige n-Gramme. Auffällig waren dabei strukturell unterschiedliche Mehrfachvorkommen ansonsten sequenziell identischer n-Gramme. Die numerischen Beiträge der einzelnen n-Gramme zum Regressionsergebnis wurden bestimmt und erstreckten sich abgesehen von einigen Ausreißern über einen Bereich von $-3{,}5$ bis $5{,}0$. Bei den vereinzelten Ausreißern handelte es sich ausschließlich um sekundärstrukturabhängige n-Gramme, deren Beiträge sich mit Werten von $-82{,}3$ und $110{,}2$ signifikant vom restlichen Wertebereich abhoben. Eine Überprüfung dieser Ausreißer im Sequenzdatensatz fand nur sehr geringe Vorkommen der reinen Sequenzen, wobei die genauen Sekundärstrukturvarianten in den Strukturensembles besonders niedrige Auftretenshäufigkeiten aufwiesen. Durch die sehr geringe Wertbelegung der dazugehörigen Deskriptoren kam es in der Regression zu einer derart hohen Gewichtung. Für die Übersicht der n-Gramm-Beteiligungen in Tabelle 6.8 wurden die n-Gramme unter Vernachlässigung ihrer Sekundärstruktur zusammengeführt. Da die Ausreißer für die tatsächliche Analyse nur geringe Bedeutung haben, wurden diese nicht in die Übersicht der Beiträge aufgenommen.

Die n-Gramm-Zusammensetzung eignet sich zwar zur Erzeugung eines mathematischen Modells, allein auf ihrer Basis können jedoch durch die fehlenden Überlappungen der sehr kurzen Fragmente keine Rückschlüsse auf größere Zusammenhänge gezogen werden. Im gegebenen Fall wurde der tatsächliche Sequenzdatensatz als Gerüst für die Aufdeckung komplexerer Zusammenhänge genutzt. Durch die Anwendung der n-Gramm-weisen Beiträge auf alle Vorkommen der n-Gramme im Sequenzdatensatz konnte die sequenzielle Zuordnung der Beteiligungen erreicht werden, die in Abbildung 6.14 gezeigt wird. Auffällig ist der höhere Grad an Strukturierung der farblich aufgetragenen Beiträge der Positivsequenzen im Vergleich zu denen der Negativse-

Tab. 6.8: Bereinigte und zusammengefasste Liste aller n-Gramme des Deskriptorensets `ngall3`, die in der Analyse des Sequenzdatensatzes der letzten Runde nach der *Feature Selection* verblieben. Im Zuge der Zusammenführung wurden die Sekundärstrukturen vernachlässigt und Ausreißer entfernt. Die Beiträge der n-Gramme zusammengeführter Sekundärstrukturvarianten wurden addiert.

n-Gramm	Beitrag	n-Gramm	Beitrag	n-Gramm	Beitrag
CTA	9.614	CTG	0.760	AGT	-1.000
TCT	3.380	CAT	0.563	TCA	-1.046
GTG	2.385	GGC	0.414	GTG	-1.130
TCG	2.281	AAA	0.336	AGA	-1.406
ATG	1.699	AGG	0.268	GAG	-1.550
GAT	1.509	TGG	0.215	TGC	-1.586
CCC	1.470	CCA	0.212	ATC	-1.674
GGG	1.401	CAA	-0.220	AGC	-1.684
GAC	1.342	GTA	-0.283	CGT	-1.826
TGT	1.334	GGC	-0.386	CAT	-1.989
ACG	1.240	CTT	-0.475	GAT	-2.361
GAG	1.130	AAC	-0.692	GGA	-2.542
GGT	1.045	TTG	-0.881	TTC	-2.683
TTC	1.003	ACG	-0.882	GAA	-3.504
GTC	0.967	CCA	-0.892	AAT	-4.031
TCA	0.964	GAC	-0.921	GCG	-4.081
CGG	0.801	CAC	-0.954	GTT	-4.330
TAT	0.779	TCG	-0.958	ATA	-5.054

quenzen. Er weist auf die tatsächliche Ordnung in den Sequenzen aus mehr oder weniger für die Bindung relevanten Fragmenten hin, die durch die Permutation bei den Negativsequenzen verloren ging. Die Gegenprobe mit permutierter Zielgröße zeigt zwar ebenfalls einen leichten Unterschied im Grad der Strukturierung zwischen den Positiv- und Negativsequenzen (hier nicht gezeigt), dieser ist jedoch nur schwach ausgeprägt und durch die deutlich höheren Modellfehler des zugrundeliegende Modells ohne Bedeutung. Im Vergleich zu den Ergebnissen der Promotorsequenzen in Abbildung 3.8 aus Kapitel 3 zeigt sich eine kaum inhomogenere Verteilung der strukturellen Elemente im Sequenzdatensatz. Während die biologischen Mechanismen für die Promotorfunktion als Ergebnis ihrer evolutionären Entwicklung nur geringe Toleranzen für die Positionierung der beteiligten Sequenzmotive lassen, können Bindestellen auf der gesamten Länge der Aptamere vorliegen, solange die Gesamtausformung der Struktur eine geeignete Formierung und Exponierung zulässt. Da die Bibliothek nur einen kleinen Teil des tatsächlichen Sequenz- und Strukturraumes abdeckte, führte diese Rahmenbedingung trotz der ansonsten großen Freiheit zu der bekannten Strukturierung aus Abbildung 6.14.

Mithilfe der nukleobasenweisen Auftragung der Beiträge zum Regressionsmodell in Abbildung 6.14 wurden zusammenhängende Subsequenzen mit positiver und negativer Bewertung bestimmt, die im Sinne komplexerer Strukturen auch über die genutzten 3-Gramme hinausgingen. Die am häufigsten anzutreffende Subsequenz mit durchgehend positiver Bewertung war das Triplett `GGT`, welches in der Regel mehrfach pro Sequenz vorkam. Sowohl das mehrfache Vorkommen als auch die aufgrund seiner Länge gering bewertete Signifikanz des Fundes führten zu dem Schluss, dass es sich bei diesem Triplett nicht um das vollständige, gesuchte Motiv handelte, wahrscheinlich jedoch um einen Teil. Selbiges traf ebenfalls auf die mehrfach pro Sequenz vorkommende, einseitige Erweiterung `GGTC` zu. Tatsächlich lagen die doppelten Vorkommen des Tripletts `GGT` häufig in großer sequenzieller Nähe zueinander, sodass in vielen Fällen die

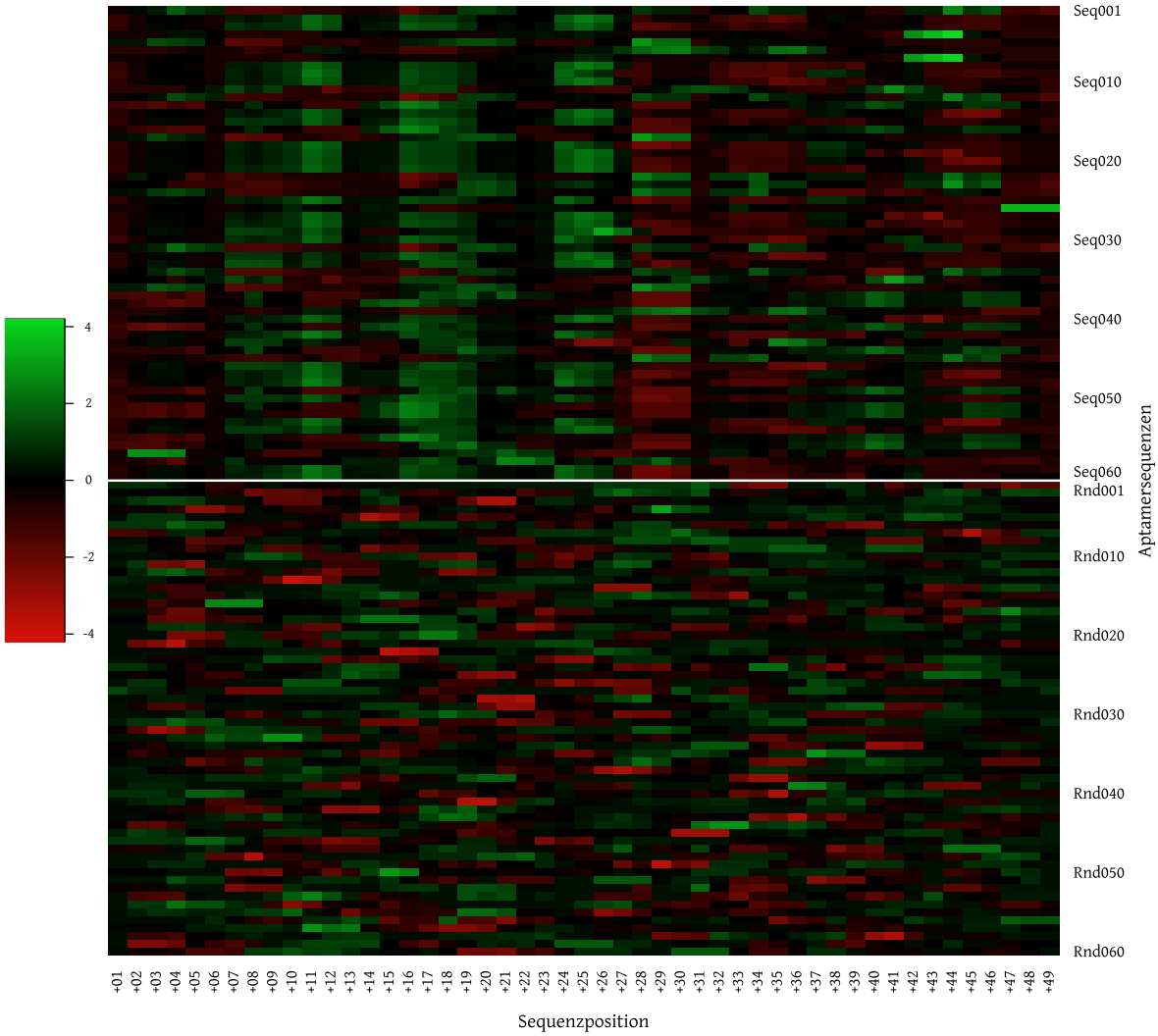


Abb. 6.14: Projektion der nukleobasenweisen Beiträge zum Regressionsergebnis auf den Sequenzdatensatz. Das PLS-Regressionsmodell wurde dafür unter Anwendung der *Feature Selection* mithilfe des Deskriptorensets `ngall3` auf dem vollständigen Datensatz trainiert, welcher Positiv- und Negativsequenzen der letzten Sequenzierrunde enthielt. Für jede Nukleobasenposition wird der Beitrag zum Regressionsergebnis in der Heatmap durch einen Farbwert dargestellt. Der Farbbereich reicht dabei von grün (positiver Beitrag) über schwarz (kein Beitrag) bis hin zu rot (negativer Beitrag). Im Bereich der Positivsequenzen fällt eine deutliche Strukturierung auf.

Sextetts `GGTGGT` positiv bewertet wurden. Wie der Blick auf die Sekundärstrukturen verriet, waren diese zu einem beträchtlichen Anteil auf *Loop*-Regionen der Aptamere positioniert. Es fanden sich jedoch auch einzelne zum Teil dominante Strukturvarianten dieses Sextetts, bei denen eine Beteiligung von Basenpaarungen die externe Zugänglichkeit einschränkte. Insgesamt bildete das Sextett `GGTGGT` eine gute Ausgangsposition für die weitere Definition eines Bindungsmusters. Unter den weiteren Funden befanden sich die Quintette `TATCA` und `CTTGC`, sowie kürzere Fragmente, die durchgehend in sehr geringer Zahl im Sequenzdatensatz vorkamen. Eine Einschätzung zu ihrer Bedeutung war auf Basis der Datenlage nicht möglich.

Entwicklung über den Experimentalverlauf Für die Bewertung der Entwicklung der Bibliothek wurden stellvertretend die Runden 3, 4 und 9 gewählt, da die erkennbare Anreicherung mit dem Übergang von Runde 3 auf 4 einsetzte und ab Runde 9 zahlenmäßig stagnierte. Die Modellfehler bei Runde 3 bestätigten, dass aus diesen Daten mit der n-Gramm-Analyse kein Nutzen

gezogen werden konnte, da noch keine hinreichende Anreicherung für die Analyse erfolgt war. Die Modellfehler der Runden 4 und 9 lagen hingegen sehr nahe an denen der finalen Runde, was die Vermutung nahe legte, dass die Aussagekraft der Modellfehler allein nicht für eine Bewertung ausreicht. Bei der Betrachtung der konkreten n-Gramm-Beiträge fielen gravierende Unterschiede auf. So konnten die extrahierten n-Gramme und deren Beiträge zum Regressionsergebnis für die Runden 3 und 4 nicht in einen sinnvollen Zusammenhang mit denen der finalen Runde gebracht werden. Für Runde 9 war zwar die Auswahl konkreter Deskriptoren durch die *Feature Selection* anders, die resultierenden nukleobasenweisen Beiträge konnten jedoch als zumindest tendenziell ähnlich mit denen der finalen Runde befunden werden. Dies gibt einen Hinweis darauf, dass die beginnende Stagnation der numerischen Anreicherung allein noch kein Garant für die vollständige Beendigung der Bibliotheksentwicklung eines SELEX-Experiments ist.

Kritische Wertung Abschließend werden die Hauptkritikpunkte des Verfahrens kurz beleuchtet. Bei der gesamten Betrachtung dieses Abschnitts blieb der Einfluss der Wahl von Negativsequenzen weitgehend unbeachtet. Zwar wurde durch die Permutation der Originalsequenzen eine sinnvolle Näherung der Hintergrundwahrscheinlichkeit erreicht, für das Ausbleiben stochastischer Effekte war der eingesetzte Sequenzdatensatz jedoch nicht umfangreich genug. Neben der Tatsache, dass die verfügbaren Daten limitiert waren, ist davon auszugehen, dass das Verfahren bei einer hinreichend starken Vergrößerung der Grunddatenmenge an seine Grenzen gestoßen wäre. Auch war bei der numerischen Ausformung des Verfahrens Vorsicht geboten. Aufgrund fehlender Affinitätsinformationen kam lediglich eine informationstheoretische Schätzung der tatsächlichen Affinitäten zum Einsatz. Außerdem basierte das Verfahren zur *Feature Selection* auf einem heuristischen Ansatz, mit dem numerisch stabile Ergebnisse nur dann mit hoher Wahrscheinlichkeit erhalten werden können, wenn passende Parameter und hinreichend tolerante Abbruchbedingungen gegeben sind. Schließlich war die fest vorgegebene Beschreibungsform der n-Gramme selbst ein Limitierungsfaktor, der keinerlei intrinsische Variabilität in den Motivfunden erlaubte. Diese konnte allenfalls im Zuge einer manuellen Nachbearbeitung der Ergebnisse gefunden werden. Zusammenfassend kann festgehalten werden, dass mithilfe der Analyse von n-Gramm-Beteiligungen wichtige Hinweise auf Bindemotive unter den Aptamersequenzen gefunden werden konnten. Aufgrund der bekannten Kritikpunkte bezüglich der Näherungsverfahren, der stochastischen Effekte und der fehlenden Unterstützung von Variabilität wurden die Sequenzdaten jedoch über ein weiteres Verfahren untersucht.

6.3.4 Durchführung einer Mustersuche

Eine flexible Form der Darstellung von Teilsequenzen wurde mit den Sequenzmustern bereits in Kapitel 4 eingeführt. Ein Muster besteht der Definition nach aus einzelnen Musterpositionen, die ihrerseits in einer gewissen Flexibilität mehrere Nukleobasen zulassen können. Über die inhaltliche Flexibilität der Musterpositionen und die Verwendung unterschiedlicher Musterlängen können Bindemotive auf der Aptamersequenz passgenauer abgebildet werden. Der ebenfalls in Kapitel 4 entwickelte Suchalgorithmus unterstützt beide Formen dieser Flexibilität. Im Gegensatz zur n-Gramm-Analyse ist das Suchverfahren für Muster mit großen Sequenzdatensätzen kompatibel ohne dabei eine Affinitätsinformation erforderlich zu machen. Es ist damit nicht anfällig für die damit verbundenen Verzerrungseffekte, was seine Anwendung auf die gegebenen Daten deutlich erleichterte. Trotzdem ist die Mustersuche der n-Gramm-Analyse bezogen auf die Aussagekraft ihrer Ergebnisse unterlegen, da neben den statistischen Kenngrößen der Muster-

funde keine Information über ihren Beitrag zur molekularen Bindung abgeleitete werden kann. Auch kann das Spannungsfeld aus positiven und negativen Beiträgen über eine Mustersuche nicht abgebildet werden. Durch die unterschiedlichen Nutzungsprofile ergänzen sich die beiden Analyseverfahren jedoch, sodass ihre Ergebnisse kombiniert wurden. Neben einem einfachen Abgleich der Ergebnisse wurden daher auch die Musterfunde dieses Abschnitts mit den bereits bestimmten Beiträgen der n-Gramm-Analyse bewertet, um aussagekräftige Schlussfolgerungen ziehen zu können.

Parametrisierung des Suchverfahrens Zu Beginn wurde das Alphabet der DNA in die Implementierung des Suchalgorithmus' integriert. Mit einer Länge von 4 ergab sich ein maximal erreichbarer Informationsgehalt von $2{,}0\,\mathrm{Bit}$ pro Musterposition. Für die Suche wurden anschließend die 1000 häufigsten Sequenzen der letzten Sequenzierungsrunde eingesetzt, sodass die Bibliothek bis auf einen definitiv zum informationellen Rauschen gehörenden Teil in die Suche einfloss. Die gesuchte Musterlänge wurde aufbauend auf den Erkenntnissen des letzten Abschnitts auf mindestens 6 festgelegt. Aufgrund der geringen Wahrscheinlichkeit sehr langer Muster konnte die obere Grenze mit 11 großzügig gewählt werden. Um die Variabilität innerhalb der Muster zugunsten ihrer Verwertbarkeit zu begrenzen, wurden pro Musterposition nur maximal zwei unterschiedliche Nukleobasen bei einem mittlerer Informationsgehalt von höchstens $1{,}8\,\mathrm{Bit}$ zugelassen. Zum Finden von Mustern mit inneren Lücken wurde zudem der Einsatz voll-variabler Musterposition als solche gestattet. Der informationelle Schwellwert zur Klassifikation einer Musterposition als Lücke wurde dabei auf $0{,}3\,\mathrm{Bit}$ festgesetzt. Schließlich wurde ein Muster nur dann in die Liste der Ergebnisse aufgenommen, wenn es in mindestens $95\,\%$ der insgesamt $233\,564$ untersuchten Einzelsequenzen gefunden wurde. Mit $90\,\%$ beim Informationsgehalt und $95\,\%$ bei der Häufigkeit waren die Schwellwerte sehr restriktiv gewählt worden, um einerseits die Menge der Suchergebnisse handhabbar zu halten und andererseits die Vergleichbarkeit der Muster zu den n-Grammen der vorigen Analyse durch moderate Variabilität zu ermöglichen. Mit den gewählten Einstellungen wurde die Mustersuche sowohl mit als auch ohne Einbezug der Sekundärstrukturinformationen durchgeführt.

Nutzung der Sekundärstrukturinformation Zwar bot sich die zweite Nutzungsvariante der Sekundärstrukturen durch ihre große informationelle Ausbeute auch für den Einsatz in der Mustersuche an, der damit verbundene algorithmische Zusatzaufwand wirkte jedoch als Ausschlusskriterium für diese Variante. Durch die Erweiterung des Alphabets wäre die notwendige Anzahl erlaubter Zeichen pro Musterposition bei gleichem Grad an Variabilität markant gestiegen, was ohne komplexe Modifikation zu einem Verlust der Kontrolle über die inhaltliche Variabilität der Musterpositionen geführt hätte. Auch skaliert das Verfahren exponentiell mit dieser Größe. Für die Mustersuche wurde daher die erste Nutzungsvariante der Sekundärstrukturinformationen gewählt, welche eine Fokussierung auf *Loop*-Bereiche bewirkt. Um die Nutzungsvariante längenunabhängig anwenden zu können, war für den Suffixbaum eine Modifikation der Erzeugungsroutine notwendig. So wurden für jede Struktur des suboptimalen Sekundärstrukturensembles alle Teilsequenzen, die keine Basenpaarung innerhalb des Aptamers ausbildeten, freigestellt und entsprechend gewichtet separat in den Suffixbaum eingefügt. Zur Handhabung des Mehrfachvorkommens einzelner Teilsequenzen wurde das Grundverfahren um einen Häufigkeitsbeiwert erweitert, der ohne Einfluss der Sekundärstruktur standardmäßig mit dem Wert 1 belegt war. Ansonsten wurde hier der relative Häufigkeitsanteil der betrachteten Sekundärstruktur im Ensemble entsprechend der Boltzmann-Verteilung eingesetzt.

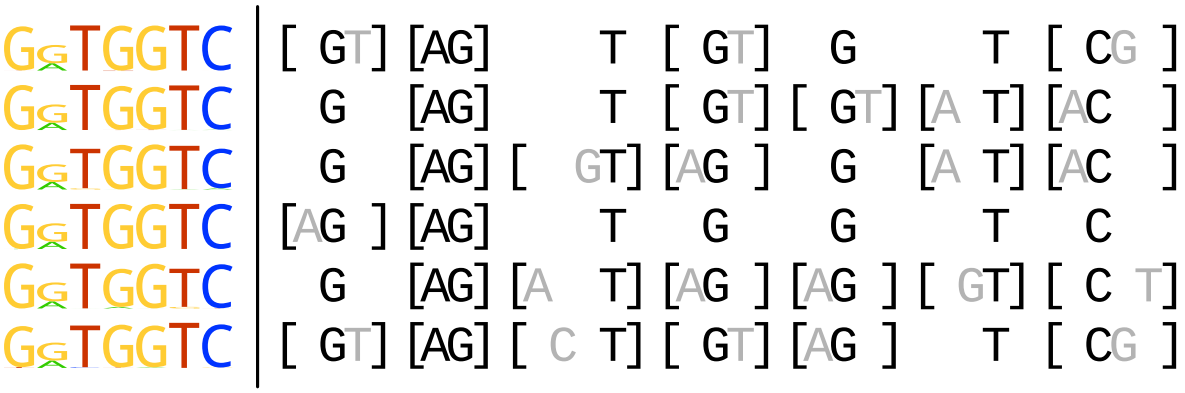


Abb. 6.15: Beispielhafte kleine Auswahl einer Reihe von Musterfunden mit großer gegenseitiger Ähnlichkeit. Die Sequenzlogodarstellungen auf der linken Seite lassen die Variabilität nur schwer bis gar nicht erkennen, da die relativen Informationen der zweithäufigsten Nukleobasen bereits so klein sind, dass die Symbole unter den häufigsten fast ganz verschwinden. Daher wurde die Darstellung der Muster als reguläre Ausdrücke in entsprechender Einrückung auf der rechten Seite ergänzt. Die im Sequenzlogo kaum oder nicht sichtbaren Nukleobasen sind im regulären Ausdruck grau dargestellt.

Ergebnisse ohne Einbezug der Sekundärstruktur Mit der entsprechenden Parametrisierung lieferte das Suchverfahren eine Menge von etwa 158 000 Einzelmotiven, welche teils sehr große gegenseitige Ähnlichkeiten zueinander aufwiesen. Ein Beispiel derartig ähnlicher Motive findet sich in Abbildung 6.15. Ursache für dieses Phänomen war die eingesetzte, erschöpfende Suchstrategie, die selbst mit der algorithmischen Erweiterung nach dem Prinzip des *Branch and Bound* noch den sinnvollen Anteil des Suchraums komplett durchmusterte. Sobald sich im Suchraum jedoch triviale Muster befinden, welche den Häufigkeitsschwellwert erfüllen oder nur knapp unterschreiten, erlaubt die eingeführte Variabilität zahlreiche kriterienkonforme Erweiterungen dieses Musters. Je geringer die Unterschreitung des Schwellwertes war, desto höher sind Anzahl und Ähnlichkeit dieser variablen Erweiterungen. Gemein ist diesen jedoch, dass sie das Hauptmotiv in der Sequenzlogodarstellung recht deutlich aufzeigen. Da die automatische Filterung der Suchergebnisse an der Formulierung eines allgemeingültigen Bewertungskriteriums scheiterte, erfolgte eine manuelle Nachbearbeitung. Der strenge Schwellwert für den informationellen Gehalt der Musterfunde vereinfachte diese jedoch, da in den meisten der betrachteten Fälle die einfach ableitbare Konsensussequenz mit dem zugrundeliegenden Hauptmotiv übereinstimmte. In den verbleibenden Fällen existierten wenige Musterpositionen mit nicht-vernachlässigbarer Variabilität, welchen durch eine schwellwertbasierte Modifikation der Konsensuserzeugung Rechnung getragen wurde. Dies erlaubte nach Abschluss des Suchvorgangs eine automatische Gruppierung aller Musterfunde entsprechend ihrer erweiterten Konsensussequenzen und reduzierte den manuellen Aufwand erheblich.

Die insgesamt 47 entstandenen Gruppen enthielten mit 36 Vertretern größtenteils lückenlose Konsensussequenzen. Die verbleibenden elf Vertreter wiesen in Bezug auf Länge und Dominanz der Lücken eine große Heterogenität auf. Abbildung 6.16 zeigt eine Übersicht über die Konsensussequenzen und deren Zusammenhänge, in der die fast durchgehend streng hierarchische Anordnung der Musterfunde bezogen auf die Enthält-Beziehung auffällig ist. Am oberen Ende dieser Hierarchie befinden sich auf der Seite der lückenlosen Motive die beiden sich stark überlappenden Sequenzen der Länge 11 `[AG]G[AG]TGGTCCGG` und `G[AG]TGGTCCGGG`, deren Vereinigung in einem separaten Suchlauf mit erhöhter Maximallänge nicht hinreichend oft aufzufinden war. Unterhalb dieser beiden Motive erstreckt sich in der Abbildung ein Netz aus Gruppen mit kürzeren Konsensussequenzen, die als Teilsequenzen implizit in der folgender Analyse einge-

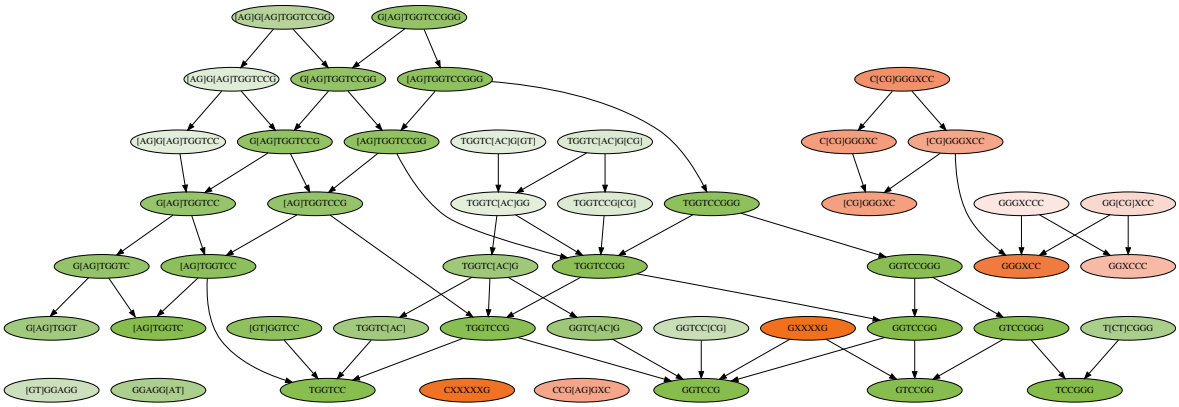


Abb. 6.16: Das Ergebnis der Mustersuche ohne Nutzung der Sekundärstrukturinformation wird als hierarchischer Graph aller gefundenen erweiterten Konsensussequenzen dargestellt. Die gerichteten Verbindungen symbolisieren die Enthält-Beziehung zweier Musterfunde, welche sowohl auf Basis der Variabilität als auch der beidseitigen Erweiterung definiert ist. Die Farbgebung zeigt an, ob ein Motiv Lücken enthält (orange) oder ob es lückenlos ist (grün). In beiden Fällen wird die Intensität des Farbtons als Indikator für die Auftretenshäufigkeit des häufigsten Vertreters der Gruppe genutzt. Die geringste Intensität (weiß) steht dabei für das minimal zugelassene Auftreten des Musters entsprechend der Parametrisierung.

schlossen wurden. Außerhalb dieses Netzwerkes lagen die zwei lückenlosen Motive `[GT]GGAGG` und `GGAGG[AT]`, die trotz relativ geringer Auftretenshäufigkeiten separat überprüft wurden. Unter den lückenhaften Funden fand sich eine kleine Hierarchie unter dem Motiv `C[CG]GGGXCC`, welches mit einer Länge von 8 zwar deutlich kürzer war als die beiden führenden Motive der lückenfreien Hierarchie, jedoch trotzdem weiter verfolgt wurde. Die verbleibenden lückenhaften Musterfunde wurden aufgrund ihrer hohen Anzahl von Lücken beziehungsweise ihres geringen Vorkommens verworfen.

Ergebnisse mit Einbezug der Sekundärstruktur In einem zweiten Suchdurchlauf wurden die Sekundärstrukturinformationen verwendet, um den Fokus der Suche auf *Loop*-Regionen zu legen. Da auf diese Weise nur noch die entsprechenden Teilsequenzen für die Suche zugänglich waren, wurde die Ergebnismenge besonders im Bereich langer Motive bei ansonsten gleicher Parametrisierung verringert. Die insgesamt 42 entstehenden Gruppen enthielten ausschließlich lückenlose Konsensussequenzen. Da jedoch die Möglichkeit bestand, dass lediglich die Lücken auf gepaarten Regionen lagen, wurden die entsprechenden Muster an dieser Stelle nicht verworfen. Die Übersicht über diese und deren Zusammenhänge ist in Abbildung 6.17 gegeben und weist ebenfalls eine klare hierarchische Strukturierung auf. Beim Vergleich mit den Suchergebnissen aus Abbildung 6.16 zeigen sich deutliche Unterschiede. Zwar gab es auch unter Einbezug der Sekundärstrukturen zwei Motive der Länge 11 an der Spitze der Hierarchie, diese überlappten sich jedoch weniger stark und gaben daher eine weniger kompakte Suchgrundlage. Während das erste Motiv fast vollständig dem Konsensusfund vor Einbringung der Sekundärstruktur entsprach, gab es für den zweiten Fund nur eine weniger vollständige Entsprechung. Infolgedessen enthielt das unter diesen beiden aufgespannte Netz aus Teilsequenzen neben einigen verwandten Treffern auch einzelne Konsensussequenzen, die nicht im vorigen Suchlauf gefunden wurden. Die Vorkommenshäufigkeiten der beiden Motive `G[AG]TGGTCCGGG` und `GGTCCGGG[GT]CC` lagen jedoch nur knapp über dem gesetzten Schwellwert, sodass die Suche nach tatsächlich relevanten Motiven im kürzeren Längenbereich geschehen konnte. Im unteren Bereich der Hierarchie bildete sich ein deutlicher Kern von Motivfunden heraus, die aufgrund ihrer hohen Auftretenshäufigkeiten besonders wahrscheinlich auf *Loop*-Regionen der Aptamere

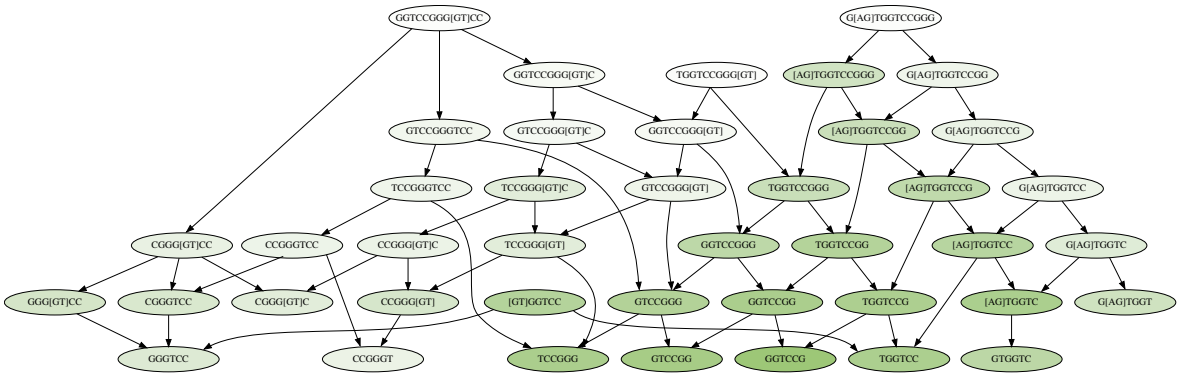


Abb. 6.17: Das Ergebnis der Mustersuche mit Nutzung der Sekundärstrukturinformation wird als hierarchischer Graph aller gefundenen erweiterten Konsensussequenzen dargestellt. Die gerichteten Verbindungen symbolisieren die Enthält-Beziehung zweier Musterfunde, welche sowohl auf Basis der Variabilität als auch der beidseitigen Erweiterung definiert ist. Das Farbschema ist aus Abbildung 6.16 übernommen.

lagen. Dieser umfasste primär den Teilbaum unterhalb der Konsensussequenz `[AG]TGGTCCGGG` mit einem Brennpunkt auf dem darin mittig gelegenen Sextett `GGTCCG`. Die folgende Betrachtung konzentrierte sich daher besonders auf diesen Bereich.

Betrachtung der Bibliotheksentwicklung Für die Betrachtung der Bibliotheksentwicklung wurden ähnlich wie in der n-Gramm-Analyse die Runden 3, 4 und 9 gewählt. Die Mustersuche in Runde 9 zeigte ähnliche Ergebnisse wie in der finalen Runde. Ohne Beachtung der Sekundärstrukturinformationen befanden sich unter ihnen bereits alle wichtigen Musterfunde der finalen Runde. Auch die Verteilung der Auftretenshäufigkeiten stimmte mit tolerierbaren Abweichungen überein. Neben den bekannten Funden wurden in dieser Runde sowohl an der Spitze als auch im unteren Bereich der Hierarchie einige zusätzliche Muster aggregiert, deren Konsensussequenzen stark variable Positionen aufwiesen. Unter Einbezug der Sekundärstruktur ergab sich ebenfalls eine breitere Hierarchie mit zusätzlichen Funden, welche den Schwellwert jedoch alle nur sehr knapp überschritten. Im Gegensatz zur letzten Runde wurde hier ein lückenhaftes Muster gefunden. Die Vorkommenshäufigkeiten der Funde waren ähnlich verteilt, jedoch insgesamt etwas geringer als in der letzten Runde. Während sich die Mustersuche in Runde 9 sehr vielversprechend erwies, ließen sich aus den Sequenzdaten der Runden 3 und 4 mangels erfolgter Anreicherung ausschließlich Muster extrahieren, die durch Lücken dominiert wurden. Eine Verwendung der Daten dieser Runden war daher ausgeschlossen.

6.3.5 Validierung der Musterfunde

Zur Validierung wurden die Ergebnisse aus der n-Gramm-Analyse und der Mustersuche ausgehend von den drei prinzipiell relevanten Musterfunden dieses Abschnitts auf gegenseitige Stimmigkeit hin untersucht. Der erste relevante Musterfund setzte sich aus den beiden überlappenden Konsensussequenzen `[AG]G[AG]TGGTCCGG` und `G[AG]TGGTCCGGG` zusammen, wobei nach Einbringung der Sekundärstrukturinformation nur noch das Sextett `GGTCCG` als Kern gefunden wurde. Nachdem der gesamte Musterfund in seine 3-Gramme zerlegt wurde, erfolgte der Abgleich dieser mit den Ergebnissen der n-Gramm-Analyse beginnend mit dem Kernmotiv. Zwar konnten durch die sehr spezifische Auswahl der *Feature Selection* nicht alle 3-Gramme des Kernmotivs in der Ergebnisliste gefunden werden, die gefundenen wiesen jedoch durchweg positive Beiträge auf. Auch für die beidseitigen Erweiterungen des Kernmotivs konnten in der

n-Gramm-Analyse einige Übereinstimmungen mit positiven Beiträgen gefunden werden, aus denen für die variablen Positionen des linken Erweiterungsbereichs eine Empfehlung für die Nukleobase Guanin hervorging. Die Güte der Erweiterungen konnte anhand der Werte jedoch nicht verglichen werden. In der nukleobasenweisen Projektion der n-Gramm-Beiträge konnten im sequenziellen Kontext auch längere Teilsequenzen des Musterfundes untersucht werden. Hier traten die überlappenden Teilsequenzen `GGTGGT`, `TGGT` und `GGTC` mit häufigem Auftreten und durchweg positiver Beitragsannotation hervor. Der zweite relevante Musterfund setzte sich aus den beiden überlappenden Konsensussequenzen `[GT]GGAGG` und `GGAGG[AT]` zusammen. Die an diesem Musterfund beteiligten n-Gramme zeigten in der Regression sehr unterschiedliche Beiträge mit einer insgesamt nur leicht positiven Tendenz. Gemeinsam betrachtet mit der relativ niedrigeren Vorkommenshäufigkeit dieses Musterfundes und der fehlenden Entsprechung in den nukleobasenweise aufgetragenen Beiträgen, wurde der zweite Musterfund vorerst als nicht relevant für die Aptamer-Bindung verworfen. Für den dritten relevante Musterfund `C[CG]GGGXCC` konnten nur im vorderen Bereich positiv bewertete n-Gramme gefunden werden. Im hinteren Bereich fand sich in der n-Gramm-Bewertung durch die Lücke ein sehr diverses Bild. Der dritte Musterfund wurde basierend auf dieser unklaren Bewertung und der Tatsache, dass das Motiv nach der sekundärstrukturbasierten Filterung nicht mehr gefunden wurde, verworfen.

Ein weiterer Aspekt der Validierung war die Überprüfung der strukturellen Lage der verbliebenen Muster in den dominierenden Strukturvarianten der zehn häufigsten Aptamerkandidaten der finalen Bibliothek. In vier der zehn Fälle war das Teilmotiv `TGGTCCGG` bereits in der optimalen Strukturvariante des Aptamers vollständig auf einer *Loop*-Region gelegen, aber auch in den verbliebenen Fällen fand sich mindestens eine passende suboptimale Strukturvariante. Für das kürzere Motiv `GGAGG` konnten hingegen für den Großteil der Aptamerkandidaten keine Strukturvarianten mit eine vollständigen Lage auf *Loop*-Regionen gefunden werden. Lediglich der hintere Teil des Motivs war in einigen suboptimalen Strukturvarianten Teil eines *Loops*. Dieser schloss jedoch relativ direkt an das vorherig betrachtete Motiv an, wobei nur in einigen Fällen ein zusätzliches Adenin zwischen den beiden Motiven lag. Das kürzere Motiv gehörte damit effektiv zur vorderen Erweiterung des Gesamtmotivs `G[AG]TGGTCCGGG`. Beispielhaft wird dies in den Strukturvarianten der zwei häufigsten Sequenzen in Abbildung 6.18 deutlich. Für die häufigste Sequenz existierten zwar zahlreiche suboptimale Strukturen, bezogen auf das zu überprüfende Motiv gliederten diese sich jedoch in zwei Hauptstrukturvarianten ein. Die erste Variante (Abbildung 6.18, links) wies eine große *Loop*-Region vor, die zwar das Motiv `GGAGG` enthielt, jedoch nur den Beginn des Motivs `TGGTCCGG`. Der *Loop* der zweiten Variante (Abbildung 6.18, mitte) trug sowohl den hinteren Teil des kurzen Motivs `GGAGG` als auch das vollständige Motiv `TGGTCCGG`. Auch bei einigen der weiteren suboptimalen Strukturvarianten konnte ein ähnlich gestalteter *Loop* beobachtet werden. Die energetische Konstellation dieser Varianten erlaubte die realistische Annahme, einen solchen *Loop* tatsächlich anzutreffen. Für die zweithäufigste Sequenz ergab sich in der betrachteten Region nur eine Strukturvariante, welche ein sehr ähnliches Bild zeigte (Abbildung 6.18, rechts).

Während der Analyse hat sich herausgestellt, dass die Sequenzierung mittels NGS-Technologie eine nützliche Grundlage für die bioinformatische Analyse gelegt hat und damit im Rahmen der Aptamerselektion als wichtiges Werkzeug zu werten ist. Die hohe Auflösung und Abdeckung zeigten ihren Nutzen besonders in den Möglichkeiten, die Entwicklung des experimentellen Verlaufs beobachten sowie statistisch mit den Sequenzdaten arbeiten zu können. Mithilfe der n-Gramm-Analyse und Mustersuche konnten die Sequenzierungsdaten unter verschiedenen Gesichtspunkten und unter Einbezug unterschiedlicher Zusatzinformationen analysiert wer-

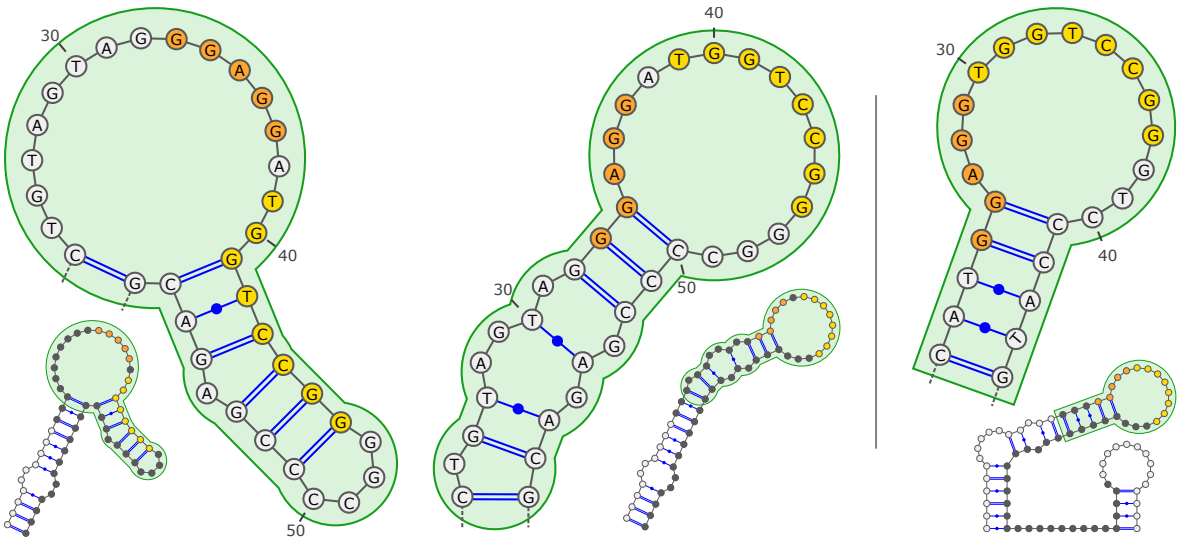


Abb. 6.18: Auftragung der gefundenen Motive auf Sekundärstrukturvarianten der häufigsten vertretenen Aptamerkandidaten. Gezeigt sind für die häufigste Sequenz der finalen SELEX-Runde die zwei wahrscheinlichsten Sekundärstrukturen, welche gleichzeitig Vertreter der beiden generellen Strukturvarianten dieser Sequenz sind (links, mitte) und die zweithäufigste Sequenz, welche nur eine Variante des relevanten Bereichs aufweist, in ihrer optimalen Struktur (rechts). Die Visualisierung erfolgte mit der Software VARNA [34]. Die gesamte Struktur ist zusätzlich in Miniatur dargestellt, wobei nicht-relevante Nukleobasen der Primer ausgelassen sind. Relevante Nukleobasen der Primer werden hellgrau dargestellt, während die der Insertregion dunkelgrau gezeichnet sind. Die vergrößerte Detailansicht ist grün hinterlegt und enthält die beiden hervorgehobenen Muster `TGGTCCGG` (gelb) und `GGAGG` (orange). Besonders der hintere Teil des Musters `GGAGG` (orange) sowie das in vielen Fällen direkt anschließende Muster `TGGTCCGG` (gelb) sind in mindestens einer der Variationen auf einer *Loop*-Region der Struktur gelegen.

den. Mit jeweils spezifischem Fokus ergänzten sich dabei die eingesetzten Analyseverfahren, sodass markante Teilsequenzen im Datensatz herausgestellt werden konnten, die mit hoher Wahrscheinlichkeit an der Bindung des Aptamers beteiligt waren. Die weitere Analyse unter Einbezug der konkreten dreidimensionalen Strukturen war jedoch unerlässlich, um die tatsächliche Geometrie der Bindung zu beleuchten.

6.4 Bioinformatische Analyse auf Basis der Tertiärstruktur

Der zweite Hauptpfad der bioinformatischen Analyse wird entsprechend dem entworfenen Verfahrensprotokoll in diesem Unterkapitel behandelt. Als informationelle Grundlage dienten die Sequenzen des Zielproteins und des am häufigsten in der finalen Bibliothek vorkommenden Aptamers. Im Zuge der Tertiärstrukturvorhersage sind ferner die geometrischen Informationen ähnlicher Strukturen aus der PDB in den Prozess eingeflossen. Über die mehrstufige Kombination aus Dockingsimulation und der Analyse der erhaltenen Komplexe wurden dabei unter steter Rückkopplung mit den Ergebnissen des vorigen Unterkapitels konkrete Bindungsgeometrien aus Aptamer und Zielprotein abgeleitet. Die methodischen Vorüberlegungen aus Kapitel 5 dienten dabei als Grundlage und Leitfaden für die Analyse der intermediären Komplexe. Die Untersuchung der gefundenen Bindestellen erlaubte mit Blick auf die Geometrie der Komplexstruktur im Kontext des gesamten Viruskapsids eine detaillierte Bewertung bezogen auf den praktischen Einsatz des Aptamers zur Detektion von Noroviren.

6.4.1 Bestimmung der Struktur des Zielproteins

Als Zielmolekül für die Aptamerselektion und damit auch für die Simulation des Aptamer-Target-Komplexes diente das große Kapsidprotein VP1 des Norovirus Genotyp GII.4 der Clustergruppe Farmington Hills, welches in Abschnitt 6.1 eingeführt wurde. Die Sequenz des 542 Aminosäuren langen Proteins war bekannt und in Tabelle 6.1 gegeben. Da sowohl in den internen Datenbeständen als auch in der Datenbank PDB keine experimentell bestimmte Tertiärstruktur für das Zielprotein existierte, wurde eine Tertiärstrukturvorhersage notwendig. Diese wurde durch eine Reihe von Templatestrukturen aus dem öffentlichen Datenbestand gespeist, bei denen es sich hauptsächlich um einzelne P-Domänen sowie ganze Kapsidproteine verwandter Norovirusstämme handelte.

Verfügbare Methoden zur Tertiärstrukturvorhersage von Proteinen Die korrekte Proteinfaltung stellt seit über 50 Jahren ein nicht zufriedenstellend gelöstes Problem in der Bioinformatik dar, zu welchem die Tertiärstrukturvorhersage gehört. Der massive Grad parallelen Ablaufes während einer Proteinfaltung lässt sich bisher nur für sehr kleine Proteine annähernd *in silico* nachvollziehen. Die stetige Zunahme experimentell aufgelöster Proteinstrukturen eröffnet jedoch im Rahmen heuristischer Methoden neue Möglichkeiten der Strukturvorhersage [505]. Abhängig von der Verfügbarkeit bereits aufgelöster und geeigneter struktureller Vorlagen bieten sich zwei prinzipielle Wege der Strukturvorhersage, welche beide mit nachgestellten Verfeinerungsverfahren noch weiter in ihrer Qualität verbessert werden können. Studien der letzten Jahre zeigten jedoch, dass eine Verbesserung der Strukturen durch derartige Verfeinerungsverfahren stets auf Teilaspekte begrenzt blieb [506; 507].

Existieren homologe Strukturen, so kann über den Vergleich der Sequenzen von sequenziellen auf strukturelle Ähnlichkeiten geschlossen werden. In der klassischen Homologiemodellierung wird hierzu eine Templatestruktur bestimmt und von dieser ausgehend mit besonderem Fokus auf den Unterschieden das Protein modelliert. Die *Coverage* gibt dabei an, welcher Anteil der zu modellierenden Sequenz durch ein Template abgebildet wird. Nur mit einer hohen *Coverage* kann ein vollständiges Homologiemodell generiert werden. Hingegen können bereits ab einer Sequenzidentität von $>50\,\%$ verlässliche Tertiärstrukturen vorhergesagt werden [508; 509]. Im Bereich geringer Ähnlichkeit greift ein weiteres Prinzip, da trotz der enormen Vielfältigkeit der Proteinfaltung häufig wiederkehrenden Teilstrukturen vorkommen. Davon ausgegangen, dass die Menge dieser Grundfaltungen von Natur aus deutlich stärker begrenzt ist, können auch in dieser lokalen Ebene homologe Teilstrukturen zur Modellierung des Gesamtproteins herangezogen werden. Diese sogenannten Threading-Verfahren nutzen damit eine wesentlich breitere Datenbasis, was sie in die Lage versetzt, bereits von sehr geringen und lokal begrenzten Ähnlichkeiten zu profitieren [510–512]. Der Umfang des Gesamtprozesses, bestehend aus Template-Auswahl, Alignment und verschiedenen Modellierungsschritten, führt in der Regel zu komplexen Pipelines, die sich zahlreicher weiterer heuristischer Methoden bedienen [513–515].

Können keine hinreichend ähnlichen Templatestrukturen gefunden werden oder sind diese aufgrund fehlender Unterscheidbarkeit ungeeignet für die Modellierung, so verbleibt die Vorhersage der Tertiärstruktur *ab initio*. Um den großen Faltungsraum der Proteine einzuschränken und damit die auf näherungsweisen physikalischen Prinzipien simulierte Faltung mit heutiger Rechentechnik zu ermöglichen, werden mit geeigneten Verfahren sowohl entfernte als auch nahe atomare Kontakte als Nebenbedingungen abgeleitet [513]. Neben normalen atomaren Kontakten [516] umfasst dies besonders die Interaktionen der bekannten Sekundärstrukturelemente [517; 518]. Durch diese implizite Nutzung der strukturellen Informationen anderer Proteine

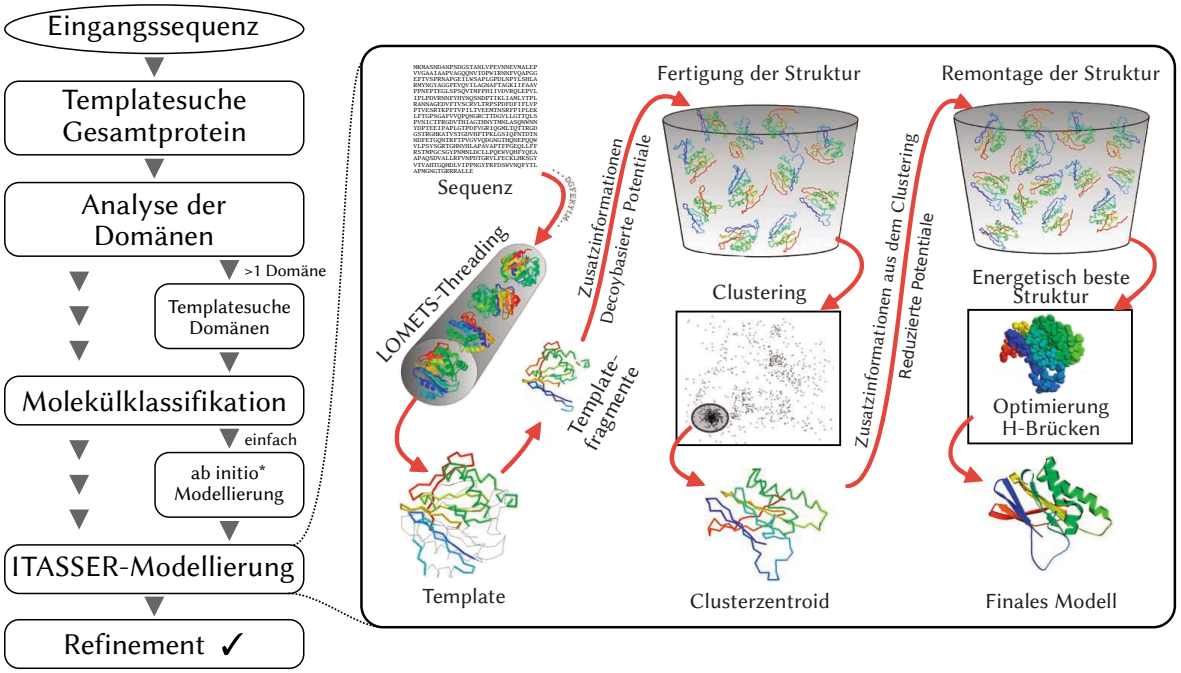


Abb. 6.19: Die Zhang-Pipeline [524] zur Homologiemodellierung von Proteinen wurde in CASP erfolgreich evaluiert. Sie kombiniert sowohl gänzliche als auch domänenspezifische Ansätze der Homologiemodellierung mit ausgereiften Verfahren der Suche und Bewertung von Template-Strukturen. Für kleine Domänen wird ferner eine Template-unterstützte *ab initio*-Vorhersage in den Modellierungsprozess mit einbezogen. Das zusammengesetzte Kernstück der Pipeline I-TASSER ist im Ablaufschema feingliedriger aufgeschlüsselt. Die Abbildung wurde modifiziert nach [524; 525].

wird eine deutliche Verbesserung des Ergebnisses erreicht [513; 519]. Zwar ist prinzipiell auch eine rein molekulardynamische Simulation möglich, um den Faltungsprozess nachzubilden, der Rechenaufwand übersteigt jedoch selbst bei Einsatz von Spezialhardware mit steigender Systemgröße und Simulationsdauer schnell die zur Verfügung stehenden Kapazitäten. Da das Simulationssystem sowohl die lineare als auch die gefaltete Form des solvatisierten Proteins beinhalten muss und die Faltungsprozesse im mehrstelligen Mikrosekundenbereich ablaufen, sind die Grenzen der aktuellen Möglichkeiten rasch erreicht [513; 520; 521].

Angewendete Strategie zur Tertiärstrukturvorhersage Um einen Überblick über die Leistung der vielen angebotenen Softwareprodukte zu erhalten, wurde der etwa zweijährlich stattfindende Wettbewerb *Critical Assessment of Techniques for Protein Structure Prediction* (CASP) ins Leben gerufen, bei dem sich Forschergruppen mit entsprechenden Werkzeugen an der Vorhersage von Proteinstrukturen unterschiedlicher Komplexität messen. Die Softwarelösungen der Zhang-Gruppe erreichten im aktuellen CASP mit einer Kombination aus gänzlicher und domänenweiser Modellierung, der Korrelation mit einer zusätzlichen Template-unterstützten *ab initio*-Modellierung sowie einem nachgeschalteten Verfeinerungsschritt hervorragende Platzierungen [522–524]. Abbildung 6.19 gibt eine schematische Übersicht dieses Verfahrens, dessen Kern durch die Software *Iterative Threading Assembly Refinement* (I-TASSER) gebildet wird. I-TASSER selbst ist dabei kein monolithisches Produkt, sondern ebenfalls eine komplexe Kombination verschiedener Softwarewerkzeuge [524]. Eine exakte Realisierung der Zhang-Server-Pipeline war jedoch im Rahmen dieser Arbeit nicht möglich, da nicht alle eingesetzten Softwa-

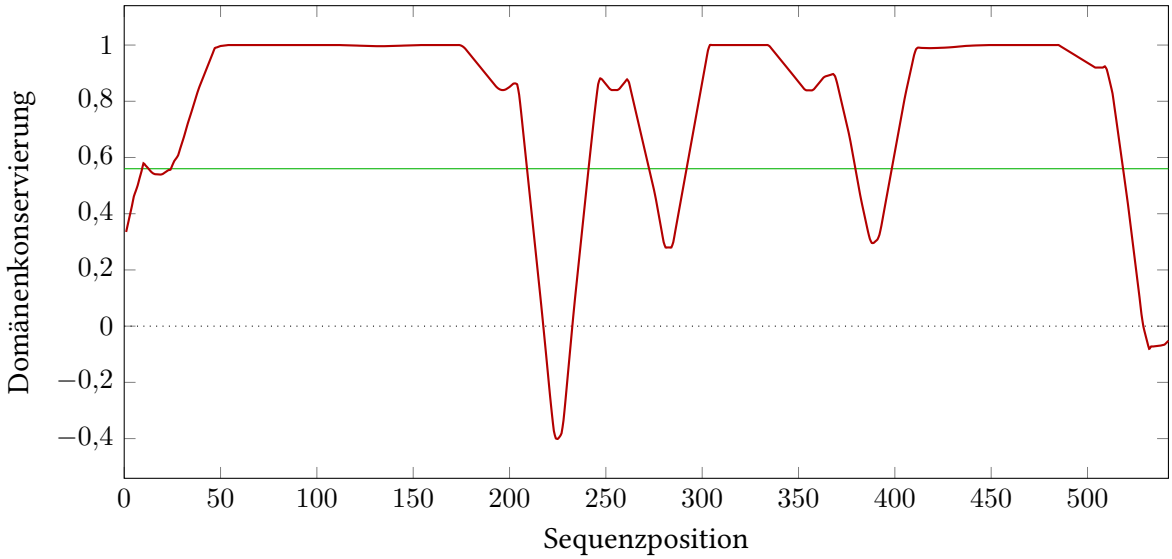


Abb. 6.20: Der Verlauf der Domänenkonservierung wurde durch die Software ThreaDom [528] basierend auf den gefundenen Templatestrukturen über die gesamte Länge der Sequenz nukleobasenweise bestimmt. Der dabei für die Unterscheidung der Domänen genutzte Schwellwert von 0,56 ist als grüne horizontale Linie eingezeichnet. Es ergaben sich die vier Domänen aus Tabelle 6.9, von denen zwei trotz sequenzieller Trennung als zusammengehörig identifiziert wurden.

rewerkzeuge als Webserver- oder Standalone-Version verfügbar waren. Die Herangehensweise bei der Vorhersage der Tertiärstruktur des Norovirus-Kapsidproteins orientierte sich jedoch stark an dieser Pipeline.

In einem ersten Schritt wurden homologe Strukturen für die gesamte Proteinsequenz über den Webserver *Local Meta-Threading-Server* (LOMETS) [526] bestimmt. Der Server lieferte dabei auf Basis von 14 unterschiedlichen Auswahlverfahren jeweils zehn nach ihrer Güte bewertete Templatestrukturen. In der manuellen Sichtung dieser Strukturen zeigten sich zwei relevante Gruppen. Während in der ersten Gruppe Kapsidproteinstrukturen unterschiedlicher Caliciviren eine hohe *Coverage* (>90 %) verbunden mit einer geringen Sequenzidentität (<50 %) aufwiesen, sammelten sich in der zweiten Gruppe hauptsächlich Strukturen von P-Domänen unterschiedlicher Norovirusstämme mit hoher Sequenzidentität (>80 %) dafür aber deutlich geringerer *Coverage* (<60 %). Da die Homologiemodellierung auf Informationen über die ganze Länge der Sequenz angewiesen ist und dabei vergleichsweise geringe Ansprüche an den Grad der sequenziellen Ähnlichkeit stellt, wurden die Strukturen der ersten Gruppe in der gemeinsamen Vergleichswertung präferiert. Unter den Kandidaten mit den besten internen Bewertungen fand sich bei fast allen Verfahren die Struktur mit der PDB-ID `1IHM` (Ketten `A` und `B`) [458], welche bereits in Abbildung 6.3a gezeigt wurde. Als tatsächlicher Vertreter der Noroviren erreichte diese eine *Coverage* von 90,4 % bei einer Sequenzidentität von 45 %. Zwar befand sich unter den Ergebnissen dank seiner höheren *Coverage* noch eine Struktur des Kapsidproteins des *Rabbit Hemorrhagic Disease Virus* (RHDV) [527], diese wurde jedoch aufgrund ihrer niedrigen Sequenzidentität nicht weiter verwendet. Anhand der von LOMETS gefundenen Templatestrukturen und deren Alignments mit der Zielsequenz bewertete die Software *Threading-based Protein Domain Prediction* (ThreaDom) [528] die Konservierung der Domänen entlang der Zielsequenz. Basierend auf der Domänenkonservierung in Abbildung 6.20 wurde die Struktur schließlich in vier Einzeldomänen unterteilt, von deinen zwei als getrenntliegend aber zusammengehörig klassifiziert wurden. Diese Aufteilung stimmte mit dem Aufbau des Kapsidproteins aus drei

Tab. 6.9: Aufteilung der Zielsequenz in Domänen mit Zuordnung der korrekten Bezeichnungen, wobei die mit Stern markierten Bereiche in der Vorhersage als zusammengehörig klassifiziert wurden. Sowohl zur Gesamtsequenz als auch zu den Hauptdomänen ist neben dem besten gefundenen Template die Bewertung des zugehörigen Homologiemodells nach dem C-Score gegeben. Die Werte für TM-Score und RMSD wurden aus dieser näherungsweise bestimmt.

Nr.	Beginn	Ende	Zuordnung	Template	C-Score	TM-Score	RMSD
-	1	542	-	1IHM/A [458]	0,71	$0{,}81 \pm 0{,}09$	$5{,}9 \pm 3{,}7$
1	1	224	S-Dom.	1IHM/B [458]	0,77	$0{,}82 \pm 0{,}09$	$4{,}0 \pm 2{,}7$
2*	225	281	P1-Dom., sub 1	3SLD/C [530]	1,29	$0{,}89 \pm 0{,}07$	$3{,}7 \pm 2{,}5$
3	282	388	P2-Dom.				
4*	389	542	P1-Dom., sub 2				

Domänen überein. Zur konkreten Zuordnung der als Domänen identifizierten Sequenzbereiche zu den namentlichen Domänen des Kapsidproteins wurden die annotierten Sequenzinformationen des Templates mit der *Accession ID* Q83884 aus der Datenbank UniProtKB [529] zurate gezogen. Tabelle 6.9 gibt einen Überblick über die Zuordnung. Die Suche nach passenden Templates wurde anschließend für die einzelnen Domänen wiederholt. Dabei ergaben sich für die S-Domäne die Struktur 1IHM (Kette B) und für die P-Domäne die Struktur 3SLD (Kette C) als bestes Template. Die Komplexität der einzelnen Domänen [514] wurde dabei in beiden Fällen als einfach eingestuft. Sie wären damit für die Template-unterstützte *ab initio*-Modellierung geeignet gewesen, die jedoch aufgrund fehlender Zugänglichkeit der Software in dieser Arbeit nicht durchgeführt werden konnte.

Sowohl die Einzeldomänen als auch die Gesamtsequenz wurden gemeinsam mit den aggregierten Zusatzinformationen über Templatestrukturen, deren Alignments und die Domänenaufteilung an den I-TASSER-Server [525; 531–533] zur Modellierung übergeben. Der Server lieferte zu jeder Eingangssequenz jeweils fünf Homologiemodelle auf Basis verschiedener Templatestrukturen. Aus diesen Modellen stach in der beiliegenden Bewertung in allen Fällen dasjenige Homologiemodell positiv hervor, welches zu der im Vorhinein als optimal bestimmten Template-Struktur gehörte. Die interne Bewertung nach dem C-Score zeigte mit Werten $>0{,}7$ in einem erlaubten Wertebereich von $[-5{,}2]$ eine gute Eignung dieser Homologiemodelle an. Die ebenfalls in Tabelle 6.9 gezeigten Abschätzungen von *Template modeling* (TM)-Score [534] und RMSD bestätigen diese. Im Vergleich zwischen domänenweiser und gänzlicher Modellierung gaben die mitgelieferten Bewertungen einen klaren Hinweis darauf, dass die separate Modellierung der beiden Hauptdomänen zu einem Anstieg der Modellgenauigkeit geführt hat. Mithilfe der lokalen Qualitätsbewertung nach ResQ [535] konnte dieser quantitative Unterschied auch qualitativ erschlossen und auf die Sequenz projiziert werden.

Wie aus Abbildung 6.21 hervorgeht, waren besonders im Bereich der P-Domäne deutliche Verbesserungen in der Modellqualität festzustellen. Während beim Gesamt-Template eine niedrigere Sequenzidentität zum Erlangen der hohen *Coverage* in Kauf genommen werden musste, konnte für die einzelne P-Domäne ein Template gefunden werden, welches sowohl eine sehr hohe Sequenzidentität als auch eine vollständige *Coverage* aufwies. Die Auswirkung auf das Modellierungsergebnis war daher durchweg positiv. Für die einzeln betrachtete S-Domäne fand sich in den Datenbeständen keine Template-Struktur mit ähnlich hohem Verbesserungspotential. Durch die allgemein höhere Modellierungsqualität in diesem Bereich war dies trotz im Mittel nur

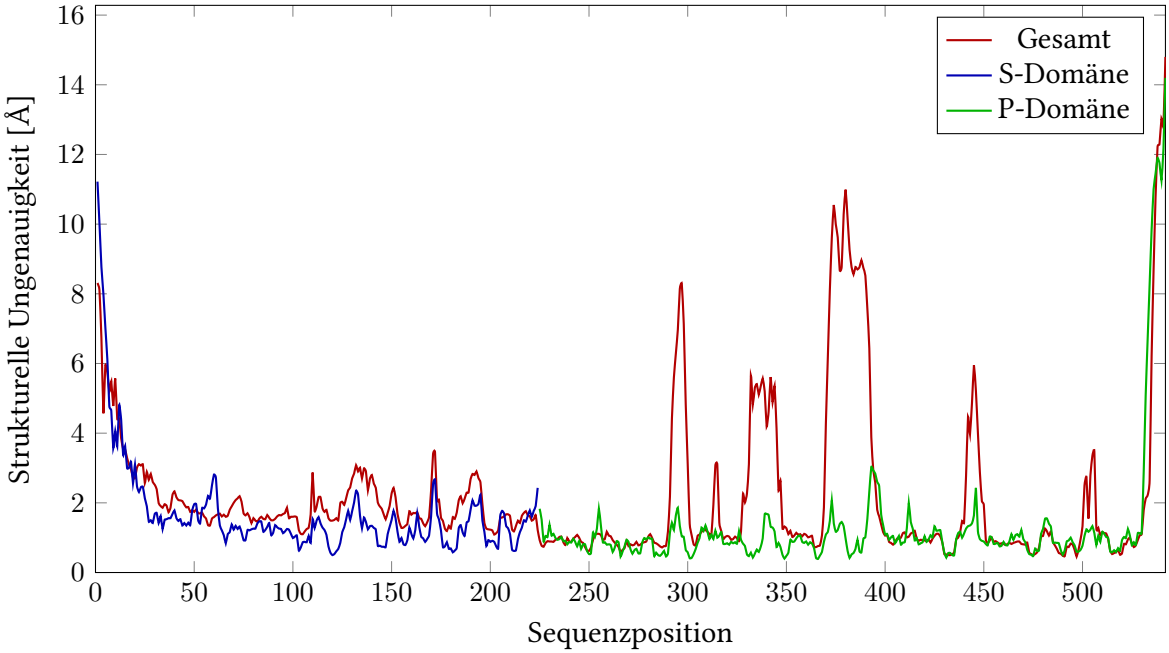


Abb. 6.21: Gezeigt wird die lokale Vorhersagegenauigkeit bei der Homologiemodellierung des Zielproteins, welche mit dem Verfahren ResQ [535] abgeschätzt wurde. Zwischen der Modellierung des Gesamtproteins (rot) und der separaten Modellierung der einzelnen Domänen (S-Domäne blau, P-Domäne grün) ergeben sich dabei qualitative Unterschiede, welche in der Differenz der entsprechenden Graphen sichtbar werden. Im vorliegenden Fall ergibt sich daraus besonders im Bereich der P-Domäne ab Sequenzposition 275 ein hohes Verbesserungspotential durch die getrennte Modellierung der Domänen.

leichter Verbesserung hier weniger kritisch. Zur Weiternutzung des erzielten Qualitätsgewinns mussten die separaten Homologiemodelle der P- und S-Domäne in eine gemeinsame Struktur vereint werden. Dazu wurden die beiden Einzelstrukturen über den Algorithmus *Combinatorial Extension* (CE)-Align [536] in eine Superposition mit der ebenfalls modellierten Gesamtstruktur gebracht. Der Algorithmus erzeugte durch die hohe sequenzielle Identität und die große strukturelle Ähnlichkeit ein verwendbares Alignment, sodass die einzelnen Domänenstrukturen unter Beachtung der Formatrichtlinien schließlich manuell zusammengefügt werden konnten. Auch wenn Domänen sowohl funktionell als auch strukturell weitgehend unabhängig voneinander sind, kommt es im Berührungsbereiche zweier Domänen mit großer Wahrscheinlichkeit zu wenigstens geringen Interaktionen. Da in dieser isolierten Betrachtung die molekularen Interaktionen zwischen den Domänen unbeachtet blieben, wurde die vereinigte Struktur einem finalen Verfeinerungsprozess durch den Server *Fragment-Guided Molecular Dynamics* (FG-MD) [537] unterzogen. Das Ergebnis der Simulation ist in Abbildung 6.22 festgehalten. Neben der hohen globalen Ähnlichkeit des Simulationsergebnisses mit der gewählten Template-Struktur sind im Vergleich (nicht dargestellt) entsprechend der Erwartung lokale Abweichungen erkennbar.

6.4.2 Bestimmung der Aptamerstruktur

Unter den zahlreichen Aptamerkandidaten wurde die am häufigsten vorkommende Sequenz der finalen Runde als Bindepartner für die Simulation des Aptamer-Zielprotein-Komplexes ausgewählt. Sie wurde bereits in Abschnitt 6.2 eingeführt und experimentell auf ihre Bindung hin untersucht. Neben dem Grundaufbau nach dem Sequenz-Template aus Tabelle 6.2 bildete die Insertsequenz aus Tabelle 6.6 den Kern dieser Aptamersequenz. Um eine vergleichende Bewer-

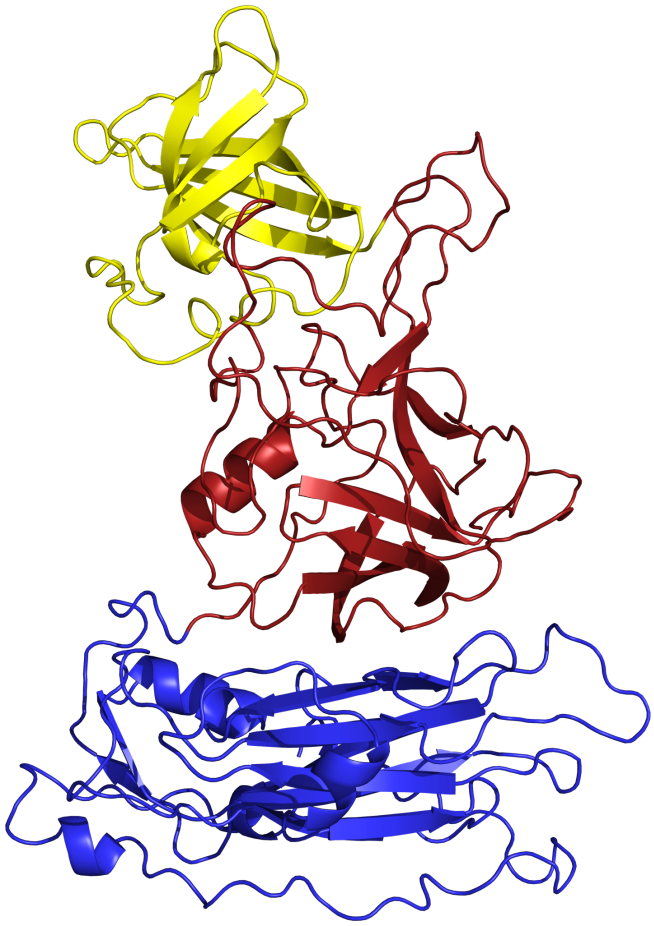


Abb. 6.22: Das Ergebnis der mehrstufigen Homologiemodellierung des Zielproteins orientiert sich in der farblichen Darstellung an der Vorgabe aus Abbildung 6.3, sodass die Unterscheidung von S-Domäne (blau) und P-Domänen (P1 in rot, P2 in gelb) erleichtert wird. Die erhaltene dreidimensionale Struktur zeigt sowohl zu den eingesetzten Template-Strukturen als auch zu den Tertiärstrukturen anderer artverwandter Kapsidproteine eine hohe globale Ähnlichkeit. In den lokalen Substrukturen finden sich jedoch hinreichend deutliche Abweichungen. Beides entspricht den Erwartungen einer erfolgreichen Homologiemodellierung.

tung über die Struktur des Aptamers und den Einfluss der Primersequenzen zu ermöglichen, wurden in der Strukturbestimmung zwei Varianten des Aptamers behandelt. Die lange Variante war mit den Primern flankiert, während die kurze Variante lediglich durch die primerfreie Insertsequenz gebildet wurde. Die Notwendigkeit einer Tertiärstrukturvorhersage ergab sich aus der fehlenden Verfügbarkeit passender Tertiärstrukturen in den internen und öffentlichen Datenbeständen.

Verfügbare Methoden zur Tertiärstrukturvorhersage von Nukleinsäuren Prinzipiell können die beiden grundlegenden Ansätze zur Tertiärstrukturvorhersage, wie sie am Beispiel der Proteine vorgestellt wurden, auch auf Nukleinsäuren angewendet werden. In der Praxis bleibt die Tertiärstrukturvorhersage von Nukleinsäuren jedoch ein Problem, dessen Lösungsstrategien weit hinter denen der Proteine zurückgeblieben sind. Die verfügbaren Lösungen erfordern häufig manuelles Eingreifen und erfahren besonders bei steigender Größe und Komplexität der Nukleinsäuren schwere Einbußen in der Vorhersagegenauigkeit. Die geringe Verfügbarkeit experimentell aufgelöster Tertiärstrukturen und die höhere Flexibilität der Nukleinsäureketten erschweren das Finden geeigneter Templatestrukturen. Dies hat zur Folge, dass auch die

homologiebasierten Methoden algorithmisch eher rudimentär umgesetzt sind [538–540]. Besonders dediziert homologiebasierte Vertreter [541; 542] bleiben daher die Ausnahme. Hinzu kommt, dass der Fokus bei der Entwicklung der Methoden fast ausschließlich auf der Vorhersage von RNA-Strukturen liegt. Im Gegensatz zu nicht-kodierenden DNA-Sequenzen ist die biologische Relevanz von nicht-kodierenden RNA-Sequenzen und daher auch deren Priorität in der Entwicklung deutlich höher.

Zur Vorhersage einer DNA-Struktur verblieben daher nur vier Methoden aus der Gruppe der *ab initio*-Verfahren. Ein recht bedenklicher Vertreter daraus stellte sich als Versuch dar, die besser entwickelten Werkzeuge der RNA-Vorhersage auf DNA-Strukturen anzuwenden. Hierbei wird die Zielsequenz in das Alphabet der RNA transformiert und anschließend auf dieser Basis eine Tertiärstruktur mit den vorhandenen Softwarewerkzeugen vorhergesagt. Durch entsprechende molekulare Modifikation wird die erhaltene Tertiärstruktur dann in eine DNA-Struktur gleichen Aufbaus zurückgeführt [543; 544]. Auch wenn die RNA-Vorhersage mit Werkzeugen wie FARFAR [545], iFOLDRNA [546; 547], 3dRNA [539] und SimRNA [548] stetig weiterentwickelt und verbessert wird, spiegelt das Ergebnis einer solchen Strategie zwangsläufig die molekularen Präferenzen der durchgeführten RNA-Faltungstrajektorie wieder. Die Nutzung der DNA-Sekundärstruktur in der Konfiguration der Vorhersage kann diesen Effekt allenfalls verringern, jedoch nicht gänzlich beseitigen. Zwei weitere Methoden wurden in Form eines Webservers angeboten, konnten jedoch im vorliegenden Fall nicht verlässlich eingesetzt werden. So unterlag der Webserver 3D-DART [549] einer strengen Limitierung auf doppelsträngige DNA-Strukturen. Bei dem Server 3dDNA handelte es sich um einen Ableger des für RNA konzipierten Vorhersagewerkzeugs 3dRNA [539], welches jedoch weder aus seiner Dokumentation noch aus einer anhängenden Publikation Informationen zur eingesetzten Modellierungsstrategie liefern konnte. Schließlich erlaubte die Simulationsplattform *Macromolecule Builder* (MMB) [550] auch die Vorhersage von DNA-Strukturen. Beginnend mit der linearen Form der DNA-Struktur wird die Faltungstrajektorie hier mit einem grobaufgelösten Modell unter Anwendung struktureller Nebenbedingungen simuliert [197; 550].

Neben den bisher aufgeführten, spezifischen Methoden kann die Tertiärstruktur einer Nukleinsäure auch über die allgemein gehaltene, aber sehr aufwändige Simulation der Molekulardynamik bestimmt werden. Hierzu ist neben der linearen Struktur des Moleküls ein passendes Kraftfeld sowie eine entsprechend komplexe Konfiguration des Simulationssystems notwendig. Die bekannten Vertreter der Kraftfelder unterstützen beide Formen der Nukleinsäuren [350]. Eine einfache Integration bekannter Sekundärstrukturinformationen durch Definition von Nebenbedingungen ist jedoch in der Molekulardynamik nicht möglich. Dies resultiert sowohl in räumlich größeren Simulationssystemen als auch in erhöhten Anforderungen an die Simulationsdauer. Das Problem des damit verbundenen, sehr hohen Rechenaufwandes wird in Form einzelner grobaufgelöster Kraftfelder adressiert. Das Kraftfeld Martini ist ein Vertreter dieser Gruppe, welcher in gängigen molekulardynamischen Simulationspaketen unterstützt wird und neben weiteren makromolekularen Typen auch Nukleinsäuren abdeckt [551; 552].

Versuch der Vorhersage ohne molekulardynamische Simulation In einer ersten Instanz wurden diejenigen Vorhersagemethoden evaluiert, die keine zeitaufwändige molekulardynamische Simulation erforderten. Der Webserver 3dDNA lieferte unter Angabe von Sequenz und Sekundärstruktur in kurzer Zeit fünf Tertiärstrukturkandidaten für das Aptamer. Neben einer Nummerierung der Dateien gab es jedoch keinerlei Hinweise auf die Güte der gelieferten Strukturen, was die weitere Verwendung erschwerte. Mithilfe von MMB [550] wurde die lineare

Strukturform der Aptamersequenz generiert und unter Zuhilfenahme der Sekundärstrukturinformation einer vereinfachten Faltungssimulation unterzogen. Die Spezifikation von Basenpaarungen und die damit verbundenen impliziten Nebenbedingungen für das *Stacking*-Verhalten der Nukleobasen beeinflussten die Faltung stark. Mit den beschriebenen Einstellungen zeigte sich in der Trajektorie jedoch ein unnatürliches Faltungsverhalten, welches durch überhöhte Packungsdichten und molekulare Kollisionen gekennzeichnet war. Selbst nach einer Simulationsdauer von $2\,\mu s$ konnte keine Konvergenz der Tertiärstruktur festgestellt werden, wobei auch die empirische Modifikation einiger kritischer Systemparameter ohne nennenswerten Effekt blieb. Der Versuch, die Simulation lediglich zur Verfeinerung der Modelle des 3dDNA-Servers einzusetzen, führte in allen Fällen zur Zerstörung der vom Server vorgegebenen Struktur mit Ausbildung der bereits beobachteten unnatürlichen Faltungscharakteristika. Im vorliegenden Fall war daher ein Zurückgreifen auf molekulardynamische Simulationen notwendig.

Vorhersage durch molekulardynamische Simulation Die Tertiärstruktur des Aptamers wurde über eine mehrstufige molekulardynamische Simulation vorhergesagt, die sowohl eine grob- als auch eine feinaufgelöste Phase beinhaltete. Als Kraftfelder wurden zur grobaufgelösten Simulation Martini [551; 552] und zur feinaufgelösten Simulation CHARMM27 [553] verwendet. In beiden Phasen kam die Simulationsplattform *Groningen Machine for Chemical Simulations* (GROMACS) [554] in Version 5.1.3 zur Anwendung. Die lineare Form des Aptamers, welche aus Kompatibilitätsgründen mit dem Tool Fiber der Suite 3DNA [198] erzeugt wurde, diente als Eingabe für die Simulation.

Die Durchführung der ersten Phase erforderte die Umwandlung der Ausgangsstruktur von der *All Atom*-Repräsentation in die reduzierte Form des Kraftfeldes Martini. In dieser reduzierten Form wurde die Struktur durch eine Reihe sogenannter *Beads* dargestellt, welche jeweils mehrere Atome der ursprünglichen Repräsentation zusammenfassten. Der komplexe Umwandlungsprozess wurde durch ein Softwarewerkzeug übernommen, welches mit dem Kraftfeld gemeinsam zur Verfügung gestellt wurde. Bedingt durch formatspezifische Inkompatibilitäten war es nicht möglich, die fünf vorhergesagten Strukturen des Webservers 3dDNA in diese reduzierte Darstellung zu überführen, sodass die Simulation nicht bereits mit einer vorgefalteten Startstruktur beginnen konnte. Das System wurde anschließend solvatisiert, indem eine entsprechend große, kubische Box um das Aptamer definiert und mit dem Solvent Wasser, ebenfalls in reduzierter Form, ausgefüllt wurde. Da das Aptamer eine negative Nettoladung trug, wurde diese im Simulationssystem durch Hinzugabe von einfach positiv geladenen Natriumionen ausgeglichen. Die Verteilung der Ionen im System erfolgte zufällig. Nach einer initialen Energieminimierung (EM) wurde das Simulationssystem in zwei unterschiedlichen Äquilibrierungsläufen sowohl auf Referenztemperatur als auch auf Referenzdruck eingestellt. Schließlich wurde die grobaufgelöste molekulardynamische Simulation durchgeführt.

In Vorbereitung auf die zweite Simulationsphase war eine Rückführung der simulierten Aptamerstruktur aus der reduzierten in die *All Atom*-Repräsentation notwendig. Das flexible, geometrische Verfahren, welches dafür genutzt wurde, positionierte die zugehörigen Atome entsprechend der *Beads* bestmöglich im Raum und bereinigte einzelne Vorkommen doppelter Atomzuordnungen automatisch. Die genauen Atompositionen ließen sich jedoch aus der reduzierten Form nicht rekonstruieren. Im Rahmen des Rückführungsprozesses wurde daher eine *All Atom*-Molekulardynamik im Bereich weniger Pikosekunden für das vorläufige Ergebnis der Rekonstruktion durchgeführt, um die Atome lokal energetisch günstig auszurichten [555]. Auf dieser Basis wurde die Topologie entsprechend des CHARMM-Kraftfeldes ermittelt und durch eine

geeignete Box umschlossen. Das System wurde anschließend unter Zuhilfenahme des elektrostatischen Dreipunkt-Wasserinteraktionsmodells TIP3P [556] solvatisiert und mit Natriumionen von seiner Nettoladung befreit. Schließlich erfolgte auch hier die Energieminimierung des Systems sowie die Einstellung der Umgebungsbedingungen in zwei Äquilibrierungsläufen, bevor die eigentliche molekulardynamische Simulation durchgeführt werden konnte.

Konfiguration der Simulation Die räumliche Nachbarschaft im Simulationssystem wurde über die Option `grid` auf benachbarte Zellen beschränkt, um eine schnellere Simulation zu ermöglichen. Zur Verwaltung der Nachbarschaftsinformationen wurde das *Cutoff*-Schema nach Verlet [557] eingesetzt. Als neuer Standard der Software GROMACS unterstützt dieses Schema die Beschleunigungstechnologien moderner Haupt- und Grafikprozessoren mit hoher Effizienz. Dies erlaubte die Parallelisierung der Berechnung zur Steigerung der Simulationsgeschwindigkeit [558]. In der ersten Phase der Simulation wurden dazu sechs handelsübliche Desktopcomputer über *Message Passing Interface* (MPI) als paralleles Rechensystem verbunden. Sie verfügten über doppelkernige Prozessoren und einfache CUDA-fähige Grafikkarten. Der hohe Mehraufwand an Rechenzeit und Netzwerkverkehr, der dabei durch die wechselseitigen Kommunikationsvorgänge zur konsistenten Simulation nötig wurde, führte jedoch zu einer nicht-idealen Skalierung der Simulationsleistung mit der eingesetzten Hardware. Für die zweite Phase der Simulation wurde daher ein Server eingesetzt, der über zwölf Prozessoren und drei High-End-Grafikkarten Tesla K40c eine effizientere interne Parallelisierung ermöglichte.

Über alle drei Dimensionen des Simulationsraumes wurden periodische Randbedingungen angelegt, sodass das Verlassen der definierten Box durch das simulierte Molekül prinzipiell kein Problem darstellte, solange es dabei nicht zu ungewollten atomaren Kontakten kam. Um derartige Kontakte zu vermeiden, überstieg die Größe der Box die des Moleküls um einen Sicherheitsbereich. Insgesamt konnte die genutzte Box durch die Anwendung der Randbedingungen deutlich verkleinert werden, sodass weniger unnötiges Solvent in der Simulation berücksichtigt werden musste. Die langstreckigen elektrostatischen Interaktionen des Systems wurden über die Methode *Particle Mesh Ewald* (PME) [559] berechnet, welche im Fourierraster mit kubischer Interpolation arbeitete. Die kurzstreckigen van-der-Waals-Interaktionen unterlagen hingegen der Behandlung eines einfachen Schwellwertsystems. In beiden Fällen wurden die Potentiale durch einen konstanten Term derart verschoben, dass sie im unteren *Cutoff*-Bereich neutral blieben.

Für die Energieminimierung wurde mit dem Integrator `steep` ein Gradientenabstiegsverfahren verwendet, während in der Äquilibrierungsphase bereits der auf *Leap Frog* [560] basierte Integrator `md` eingesetzt wurde. In beiden Verfahren wurden die makromolekularen Rahmenstrukturen derart fixiert, dass nur lokale Änderungen der Molekülgeometrie zugelassen wurden. Mit den beiden Äquilibrierungsphasen ging die schrittweise Aktivierung von Temperatur- und Druckstabilisierung einher. Die Referenztemperatur von $300\,\mathrm{K}$ wurde dazu mit einem Berendsen-Termostat [561] hergestellt, welches aufgrund seines schnellen Ansprechens für die Äquilibrierung empfohlen wird [562]. Zum Erreichen des Referenzdruckes von $1\,\mathrm{bar}$ wurde aus gleichem Grund ein Berendsen-Barostat [561] eingesetzt. Mit Beginn der eigentlichen molekulardynamischen Simulation wurde die Fixierung der makromolekularen Rahmenstrukturen gelöst. Zur Erhaltung der Referenzbedingungen konnten nun mit dem *Velocity Rescaling*-Thermostat [563] und dem Parrinello-Rahman-Barostat [564] genauere Verfahren eingesetzt werden. Die Angaben zu den Dauern der einzelnen Zeitschritte und der gesamten Simulationsphasen in Tabelle 6.10 beziehen sich auf die lange Variante des Aptamers. Die Simulation der kurzen Vari-

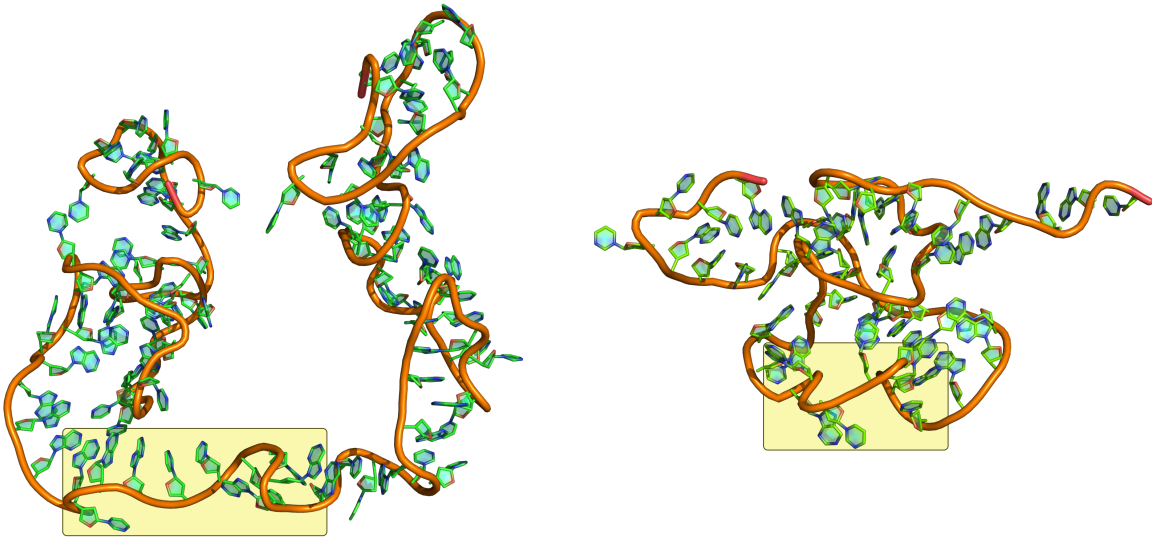
Tab. 6.10: Zeitliche und methodische Parametrisierung der grob- (CG) und feinaufgelösten (AA) Molekulardynamik in den entsprechenden Schritten der Simulation. Die angegebenen Zeiten der grobaufgelösten Simulation sind mit dem effektiven Skalierungsfaktor 4 zu verrechnen. Die Angabe der Dauer für das Gradientenabstiegsverfahren des Integrators EM bezieht sich auf die maximale Anzahl von Abstiegsschritten.

Typ	Phase	Integrator	Thermostat	Barostat	Zeitschritt	Dauer
CG	EM	steep	—	—	—	<50 000
CG	Äq.	md	Berendsen	—	1 fs	100 ps
CG	Äq.	md	Berendsen	Berendsen	10 fs	200 ps
CG	Sim.	md	*Velocity Rescaling*	Parrinello-Rahman	10 fs	500 ns
AA	EM	steep	—	—	—	<50 000
AA	Äq.	md	Berendsen	—	2 fs	100 ps
AA	Äq.	md	Berendsen	Berendsen	2 fs	100 ps
AA	Sim.	md	*Velocity Rescaling*	Parrinello-Rahman	2 fs	250 ns

ante wurde durch die frühzeitig eintretende Konvergenz der Struktur bereits nach der halben Hauptsimulationsdauer beendet. Bei der Beurteilung der angegebenen Zeiten ist zu beachten, dass die zeitliche Skalierung des grobaufgelösten Verfahrens (CG) durch die Vereinfachung des Kräftemodells effektiv um den Faktor 4 beschleunigt ist [551].

Ergebnisse der Simulation Während der Simulation wurden in fest definierten Zeitschritten sowohl eine Reihe von physikalischen Kennwerten als auch Abbilder des Faltungsfortschritts aufgezeichnet, anhand derer ihr Verlauf beurteilt werden konnte. Druck und Temperatur wurden im Mittel auf den definierten Zielwerten gehalten, auch wenn sie dabei im üblichen Rahmen liegende Fluktuationen aufwiesen. Der energetische Verlauf unterlag zwar starken Schwankungen, zeigte jedoch über die Dauer der Simulation hinweg eine sinkende Tendenz. Eine manuelle Begutachtung der aufgezeichneten Faltungstrajektorie gab ebenfalls Aufschluss über das Konvergenzverhalten der beiden Simulationssysteme. Hier zeigte sich bei der Aptamervariante mit Primersequenzen eine langsamere Konvergenz der Faltung, welche der Größe des Makromoleküls von 90 nt geschuldet war. Die resultierende Struktur dieser Variante bildete, wie in Abbildung 6.23a gezeigt, keine kompakte Faltung aus. Vielmehr formten sich zwei Arme mit jeweils interner Stabilisierung, welche über eine weniger stabile Mittelregion verbunden waren. Innerhalb dieser Mittelregion selbst fanden sich deutlich weniger stabilisierende Interaktionen. Entsprechend war auch nach der Konvergenz eine moderate Fluktuation der Struktur in der Trajektorie zu beobachten. Bei der kürzeren, primerlosen Aptamervariante konnte hingegen eine deutlich schnellere Konvergenz beobachtet werden. Während der Faltung bildete sich bei dieser Variante eine sehr kompakte Struktur aus, die nach Eintreten der Konvergenz nur noch geringe intrinsische Bewegungen auswies. Abbildung 6.23b zeigt diese Faltung.

Die Sekundärstrukturausformung der simulierten Aptamerstrukturen zeigt im Vergleich mit den vorherig auf Sequenzebene durchgeführten Vorhersagen ein eher geteiltes Bild. Die tatsächlich beobachtete Ausbildung von Basenpaarungen und *Stacking*-Interaktionen wich dabei deutlich von dem ab, was die sequenzbasierte Vorhersage vorgab. In beiden Aptamervarianten konnte jedoch auf Basis der bioinformatischen Analyseergebnisse der 2D-Ebene eine wichtige Gemeinsamkeit in den simulierten Strukturen festgestellt werden. Es handelt sich dabei



Lange Aptamervariante **(a)** **(b)** Kurze Aptamervariante

Abb. 6.23: Das Ergebnis der Tertiärstrukturvorhersage der Aptamervarianten zeigt zwei sehr unterschiedliche Faltungen durch das Hinzunehmen (a) und Weglassen (b) der eingesetzten Primersequenzen. Im zentralen Bereich befindet sich jedoch in beiden Strukturen ein identischer Sequenzbereich (gelbe Hervorhebung), der ungebundene Nukleotide aufweist.

um einen zentralen Bereich der Aptamere mit der Sequenz `TGGTCCGG`, welcher in einem Großteil der Sekundärstrukturen der vorhergesagten Ensembles nicht an Basenpaarungen teilnahm. Eben dieser Bereich konnte sowohl in der Betrachtung der n-Gram-Bewertungen als auch in der Auswertung der größer angelegten Mustersuche als relevant für die Bindung festgestellt werden. Wie Abbildung 6.23 zeigt, ist der Bereich zwar in den beiden simulierten Strukturen unterschiedlich stark gestaucht, jedoch tatsächlich kaum in die innermolekularen Interaktionen des Moleküls eingebunden. Bezogen auf die Sekundärstrukturvorhersage kann an dieser Stelle zusammengefasst werden, dass zwar in der konkreten Ausformung der Basenpaare große Abweichungen zur simulierten Struktur festgestellt wurden, die Information zur Gebundenheit einzelner Nukleotide jedoch tendenziell gut übereinstimmte. Da in den zweidimensionalen Analyseverfahren die Sekundärstruktur nur in Form dieser booleschen Gebundenheitsinformation eingeflossen war, sind die Ergebnisse dieser Analysen nicht wesentlich von den beobachteten Abweichungen der Basenpaarung beeinträchtigt. Mögliche konformationelle Anpassungen während der Bindung vorausgesetzt, bestätigt die dreidimensionale Simulation der Strukturen ferner die Konservierung des zentralen Sequenzbereiches und unterstreicht damit gleichsam dessen Bedeutung für die Aptamerbindung.

6.4.3 Bildung eines Komplexes aus Aptamer und Zielprotein

Nach der Bestimmung der Strukturen beider Komplexpartner erfolgte die Bildung einer Komplexstruktur aus Aptamer und Zielprotein durch semi-flexible Dockingsimulation. Eingesetzt wurde dazu die Software-Suite HADDOCK in Version 2.2 [352; 387], die durch entsprechende Parametrisierung sowohl Proteine als auch Nukleinsäuren als Eingabe verarbeiten kann [375]. Zwar lag das Hauptaugenmerk bei der Simulation auf der kurzen Aptamervariante ohne Primersequenzen, zur vergleichenden Betrachtung erfolgte jedoch zudem eine Simulation mit der

primerbehafteten Variante des Aptamers. Die Informationen zu dem gefundenen Motiv der Aptamersequenzen wurden in diesen Simulationslauf nicht eingebracht, um das natürliche, von der bisherigen Analyse weitgehend unbeeinflusste Bindungsverhalten beobachten zu können. Auf diese Weise können starke Präferenzen erkannt werden, es ist jedoch auch eine deutlich größere Anzahl von Komplexvarianten notwendig, um die tatsächliche Konfiguration finden zu können. Von den $10\,000$ zufällig initialisierten Kombinationen der Bindepartner wurden in der semi-flexiblen und solvatisierten Phase der mehrstufigen Simulation die 2000 am besten bewerteten Kandidaten weiterverfolgt und verfeinert.

Initiale Dockingsimulation zwischen Aptamer und Zielprotein Um das Verhalten der Dockingsimulation grob zu validieren, wurden die Bewertungen nach ITScore-PR, HADDOCK und AMBER auf Basis ihrer Verteilung unter den Ergebnisstrukturen betrachtet. Allen Bewertungsmaßen gemeinsam ist die Eigenschaft, dass kleine Werte für Strukturen mit guten Bindungseigenschaften stehen. Nach dem Ausschluss vereinzelter Ausreißer repräsentierten die Werte für die kurze Variante des Aptamers mit Streubereichen von etwa -200 bis 50 bei ITScore-PR, -175 bis 75 bei HADDOCK und -1250 bis 1250 bei AMBER sowohl energetisch unerwünschte als auch günstige Verbindungen in großer Vielfalt. Die breiten Verteilungen entsprachen der Erwartung, da die Dockingsimulation ohne Angabe von Bindepräferenzen ausgeführt wurde. Bei der Betrachtung der Histogramme in Abbildung 6.24 fällt die leichte Linksschiefe der ansonsten nahezu Normalverteilungen ins Auge. Sie ist das Ergebnis der stattgefundenen Selektion von Kandidaten zwischen den einzelnen Phasen des Dockingprozesses. Dabei wurden die ursprünglich wesentlich breiteren Verteilungen der initialen Strukturen durch eine strenge Selektion am rechten Rand hart beschnitten. Die in den folgenden Phasen der Simulation eingeführten strukturellen Freiheitsgrade resultierten durch die lokalen Veränderungen der Strukturen nicht nur bei der energetischen Bewertung in einer gewissen Unschärfe dieser Randregion. Im Gesamtbild manifestierte sich daher in allen Fällen ein steileres Gefälle am rechten Rand der Häufigkeitsverteilungen. Bei der Betrachtung der langen Variante des Aptamers wichen zwar die Grenzen der Wertebereiche etwas ab, die beschriebene Charakteristik war jedoch ebenso deutlich ausgeprägt. Zusammenfassend kann festgehalten werden, dass das Verhalten beider Dockingsimulationen auf dieser numerischen Ebene den Erwartungen weitgehend entsprach. Die jeweils 2000 erhaltenen Komplexvarianten wurden nun manuell gesichtet, um starke Präferenzen im Bindungsverhalten der beiden Komplexpartner festzustellen, die sich bereits im ersten Stadium der Simulation abzeichneten. Dazu wurden die tatsächlich auftretenden Binderegionen aller 2000 Konstellationen jeweils auf der Oberfläche des Aptamers und auf der des Proteins aufgetragen. Im Ergebnis zeigten sich weder signifikante Präferenzen noch Aussparungen für die Lage möglicher Binderegionen, sondern eine nahezu vollständige Abdeckung der jeweiligen Oberflächen mit nicht aussagekräftigen Schwankungen in ihrer Dichte. Die Inkorporation der Primersequenzen zeigte dabei keinen Einfluss.

Überprüfung auf Bindepräferenzen Da die Ergebnismengen zahlreiche nicht-relevante Komplexe mit durchgehend negativen Bewertungen enthielten, konnte die Ableitung von Bindepräferenzen nur über eine weitere Reduktion der Strukturen unter Nutzung der bekannten Bewertungsmaße geschehen. Aus diesem Grund wurden die jeweils 100 am besten bewerteten Komplexe anhand der drei Bewertungsmaße AMBER, HADDOCK und ITScore-PR freigestellt und entsprechend ihrer Wertung sortiert. Für die Komplexe zwischen Zielprotein und langer Aptamervariante zeigten sich auch hier kaum nutzbare Präferenzen auf den Oberflächen beider

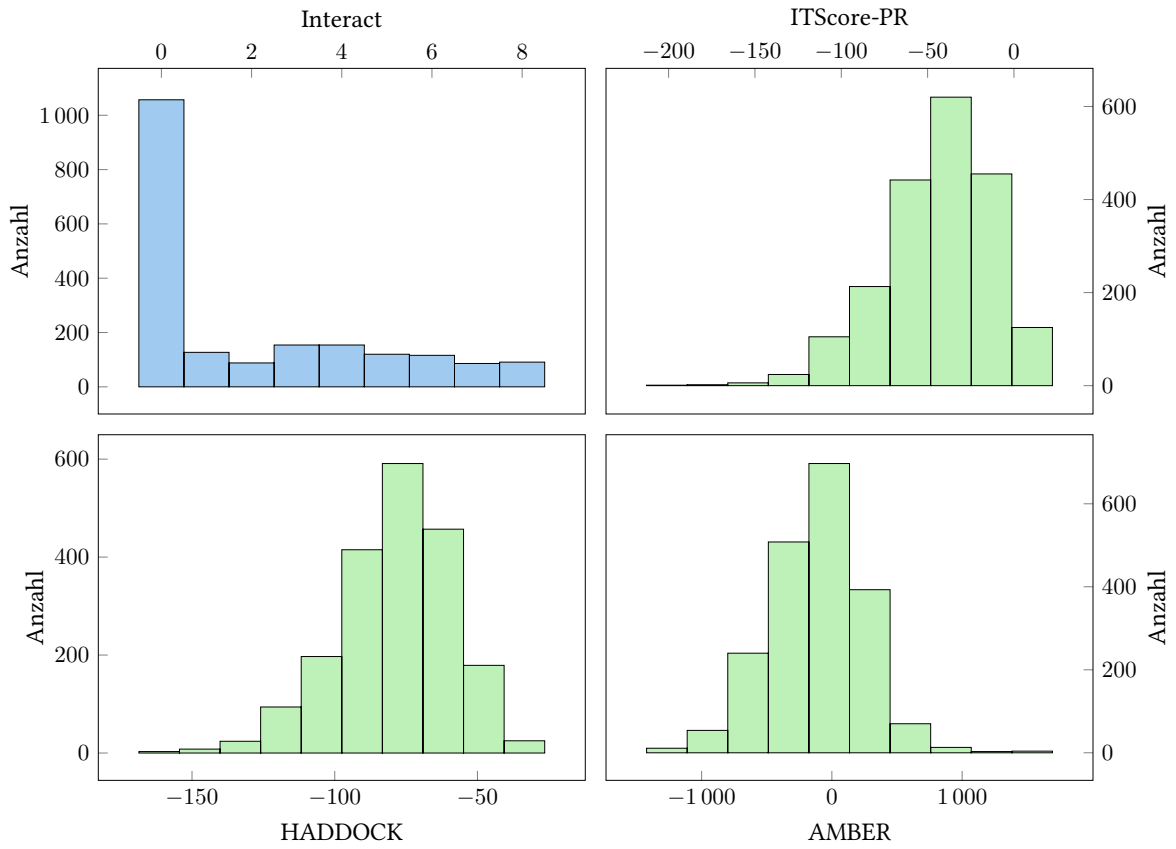


Abb. 6.24: Abgebildet sind die Häufigkeitsverteilungen vierer Kennwerte unter den 2000 Ergebniskomplexen der Dockingsimulation zwischen kurzer Aptamervariante und Zielprotein. Die drei Kennwerte ITScore-PR, HADDOCK und AMBER (grün) weisen dabei eine sehr große Streuung auf, die sowohl in den positiven als auch weit in den negativen Bewertungsbereich hineinragt und typisch für eine Simulation ohne Zusatzinformationen ist. Die leichte Linksschiefe der nahezu normalverteilten Histogramme geht auf die Selektion der am besten bewerteten Strukturen im vorderen Teil des Dockingprozesses zurück. Die Bewertung nach Interact (blau) zeigt, dass das Sequenzmotiv des eingesetzten Aptamers im Großteil der Komplexe nicht oder nur kaum an der Bindung zum Zielprotein beteiligt ist.

Partner. Lediglich bei den Bewertungen nach AMBER und ITScore-PR verblieben am oberen Ende der P2-Domäne sowie am unteren Ende der S-Domäne mit vereinzelten Ausnahmen Aussparungen auf der Proteinoberfläche. Die kompaktere Faltung der kurzen Aptamervariante begünstigte hingegen durch seine geringere Anfälligkeit für strukturell bedingte, jedoch irrelevante Wechselwirkungen die Ausbildung einiger Präferenzen. Sowohl bei der Bewertung nach AMBER als auch bei der nach ITScore-PR konnte in der detaillierten Betrachtung der am besten bewerteten Strukturen festgestellt werden, dass sich für die Aptamerbindung im wesentlichen zwei Epitope der Proteinoberfläche manifestierten. Unter den Strukturen waren diese ansonsten gleichmäßig verteilt. Das erste Epitop befand sich im oberen Bereich der P1-Domäne fast genau gegenüber des Ansatzes zur P2-Domäne. Das zweite Epitop lag hingegen an einem anderen Seitenbereich der S-Domäne mit teilweisen Ausläufern in den Bereich des Domänenübergangs von S zu P1. Als die Ergebnisse entsprechend der Bewertung nach HADDOCK sortiert wurden, kam zu diesen noch ein drittes Epitop hinzu. Dieses dritte Epitop befand sich ebenfalls im Übergangsbereich von der S- zur P1-Domäne, lag jedoch räumlich hinreichend von den beiden anderen Epitopen isoliert. Auf der Oberfläche des Aptamers ließen sich keine so differenzierten Präferenzen für die Bindung erkennen. Lediglich der obere Bereich der Struktur in der Orien-

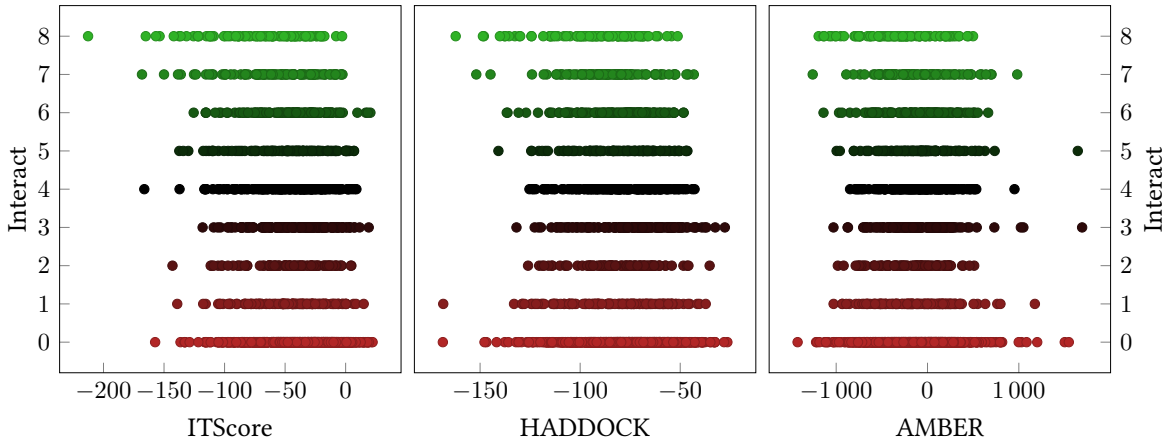


Abb. 6.25: Zur genauen Aufschlüsselung der Beteiligung des Motivs TGGTCCGG an den molekularen Schnittstellen der Aptamer-Zielprotein-Komplexe wurde die Werteverteilung der drei Bewertungsmethoden ITScore-PR, HADDOCK und AMBER getrennt nach der Kennzahl Interact aufgetragen. Grundlage für diese Untersuchung war die kurze Aptamerstruktur. Insgesamt kann dabei festgestellt werden, dass bei allen Interaktionsgraden eine breite Streuung der Bewertungen zu verzeichnen ist, welche allenfalls am negativen Ende des Wertebereiches eine leichte Tendenz erkennen lässt. Bei der Bewertung nach ITScore-PR sticht die Struktur mit der besten Bewertung deutlich heraus und gibt damit einen wichtigen Hinweis.

tierung von Abbildung 6.23b war in den Komplexen mit den besten Bewertungen nicht an der Bindung beteiligt. Da die bereits in der zweidimensionalen Analyse als bindungsrelevant gefundene Teilsequenz TGGTCCGG dieser Aussparung genau gegenüber lag, kam es bei einem großen Anteil der betrachteten Komplexe zu einer wenigstens kleinen Beteiligung dieser Teilsequenz an der Bindung.

Im Weiteren wurde daher die Bedeutung des Motivs TGGTCCGG für die Aptamerbindung durch eine genaue Analyse seiner Beteiligung an den molekularen Schnittstellen der Komplexe erschlossen. Zur Quantifizierung dieser Beteiligung wurde der Kennwert Interact eingeführt, der in den beobachteten molekularen Schnittstellen die Anzahl der Residuen angab, die zum betrachteten Motiv zugehörig waren. Zu seiner Bestimmung wurden alle paarweisen atomaren Kontakte der Schnittstelle innerhalb des Distanzschwellwertes von 8 Å ermittelt und anschließend die daran beteiligten Residuen mit dem Motiv abgeglichen. Die Häufigkeitsverteilung des Kennwertes in Abbildung 6.24 zeigt deutlich, dass der Großteil der 2000 Komplexvarianten keine oder nur eine geringe Beteiligung des Motivs an der molekularen Schnittstelle aufwies. Da im Zuge der manuellen Überprüfung unter Einbezug der Bewertungsmaße eine gegenläufige Tendenz erkennbar war, wurde für die drei verwendeten Bewertungsmaße jeweils die Korrelation zum Interact-Kennwert grafisch aufgetragen. Das Ergebnis für die kurze Aptamervariante befindet sich in Abbildung 6.25 und zeigt, dass die energetische Bewertung nach AMBER und HADDOCK unabhängig von der Beteiligung des Motivs jeweils einen großen Teil des verfügbaren Wertebereichs abgedeckt hat. Der allenfalls schwachen Tendenz einer energetisch günstigeren Bewertung von Komplexen mit hoher Beteiligung des Motivs an der Binderegion war dabei keine Signifikanz zuzumessen. Auch wenn die Bewertung nach ITScore-PR in Kapitel 5 die besten Vorhersageeigenschaften vorbrachte, zeigte auch sie in dieser Betrachtung nur die bereits benannte, schwache Tendenz. Der Komplex mit der besten Bewertung nach ITScore-PR stach jedoch in zwei Hinsichten aus der Menge der Dockingergebnisse hervor. Nicht nur setzte er sich mit einem relativ großen Vorsprung in der Bewertung von allen weiteren Komplexstrukturen ab, er wies auch eine volle Beteiligung des Motivs an der Binderegion vor. Bereits in den Untersuchungen des besagten Kapitels zeichnete sich ab, dass bei einer Dockingsimulati-

on ohne Angabe von Bindungsinformationen nur sehr wenige Vertreter einer wirklich nahenativen Konformation zu erwarten sind. In der Werteverteilung traten diese als besonders hervorstechend positive Einzelwerte hervor, was genau mit dem Bild der aktuellen Beobachtung einherging. Dem gefundenen Einzelfall wurde daher besonders in Anbetracht der bisherig übereinstimmenden Hinweise aller durchgeführter 2D- und 3D-Analysemethoden eine wichtige Bedeutung beigemessen. Das Motiv TGGTCCGG wurde schließlich als relevanter Zielbereich auf der Oberfläche des Aptamers festgehalten. Durch die Betrachtung der langen Variante des Aptamers konnten mit dieser Methodik keine Hinweise auf einen solchen relevanten Bereich der Oberfläche gefunden werden.

Formale Spezifikation der Interaktionen Für die informationsgetriebene Dockingsimulation können Bindungsinformationen in Form sogenannter AIR bereitgestellt werden, um bestimmte Bereiche des Bindungsraumes zielgerichtet zu explorieren. Über aktive und passive Residuen werden dabei Regionen beider Bindepartner als unterschiedlich stark relevant für die Ausbildung der intermolekularen Schnittstelle festgelegt. Während der Simulation werden diese vom Dockingsystem schließlich zur Einschränkung der simulierten Geometrien eingesetzt, sodass im Großteil der erhaltenen Ergebniskomplexe Bindekonstellation erhalten werden, welche die definierten Residuen einschließen. Sowohl in der genauen Positionierung als auch in der Orientierung bleiben dabei während der Simulation hinreichend große Freiheitsgrade, um eine gute Streuung im vorgegebenen Teil des Bindungsraumes zu ermöglichen.

Die aktiven Residuen sind dabei jene, die verpflichtend zur Ausbildung der gewünschten Schnittstelle notwendig sind. Ihre Definition erfolgte über die genaue Spezifikation der Kernbereiche der gefundenen Oberflächenregionen. Das Sequenzmotiv des Aptamers war über die vorgeschalteten Analysen bereits hinreichend genau bestimmt, sodass die acht bekannten Nukleotide direkt übernommen werden konnten. Um die Kernbereiche der drei Epitope des Zielproteins bestimmen zu können, wurden zuerst die 50 nach HADDOCK am besten bewerteten Strukturen manuell von all den Vertretern bereinigt, deren Binderegionen markant von den drei Epitopen abwichen. Für die verbliebenen 42 Komplexe wurde nach dem bekannten Verfahren ermittelt, welche Aminosäuren an der molekularen Bindung mit einem maximalen Atomabstand von $8\,\text{Å}$ beteiligt waren. Die Häufigkeiten dieser Vorkommen verteilten sich auf die drei Epitope und wurden pro Aminosäure kumuliert. Über einen einfachen Häufigkeitsfilter mit dem Schwellwert 7 war es nun möglich, die Kernbereiche der drei Epitope aus den Daten zu extrahieren. Die resultierenden 81 Aminosäuren wurden schließlich nach ihrer Zugehörigkeit zu den drei disjunkten Epitopen in entsprechende Gruppen eingeteilt. Zur visuellen Unterstützung wurden die ermittelten Kernbereiche in Abbildung 6.26 auf die Oberflächen der beiden Bindepartner aufgetragen. Die genaue Aufschlüsselung der aktiven Residuen befindet sich in Tabelle 6.11. Durch die gewählte Verfahrensweise der Bestimmung wurde die geforderte relative Lösungsmittelzugänglichkeit der aktiven Residuen von mindestens $50\,\%$ [375] sichergestellt.

Die passiven Residuen sind zwar zur Ausbildung der molekularen Schnittstelle nicht zwingend erforderlich, sie bilden jedoch einen strukturellen Toleranzbereich, der dem Dockingsystem die notwendige Flexibilität erlaubt. Auf diese Weise wird ein Ausgleich für auftretende Fehler und Ungenauigkeiten in der Bestimmung der Zielepitope geschaffen und sichergestellt, dass eine hinreichend große Vielfalt unter den Ergebnissen zur Beurteilung des Bindungsraumes beitragen kann. Entsprechend der Dokumentation des HADDOCK-Systems wurden alle hinreichend lösungsmittelzugänglichen Residuen in der direkten Nachbarschaft des Kernbereiches als passiv angenommen. Ihre Bestimmung erfolgte manuell und nach Epitopen getrennt über den

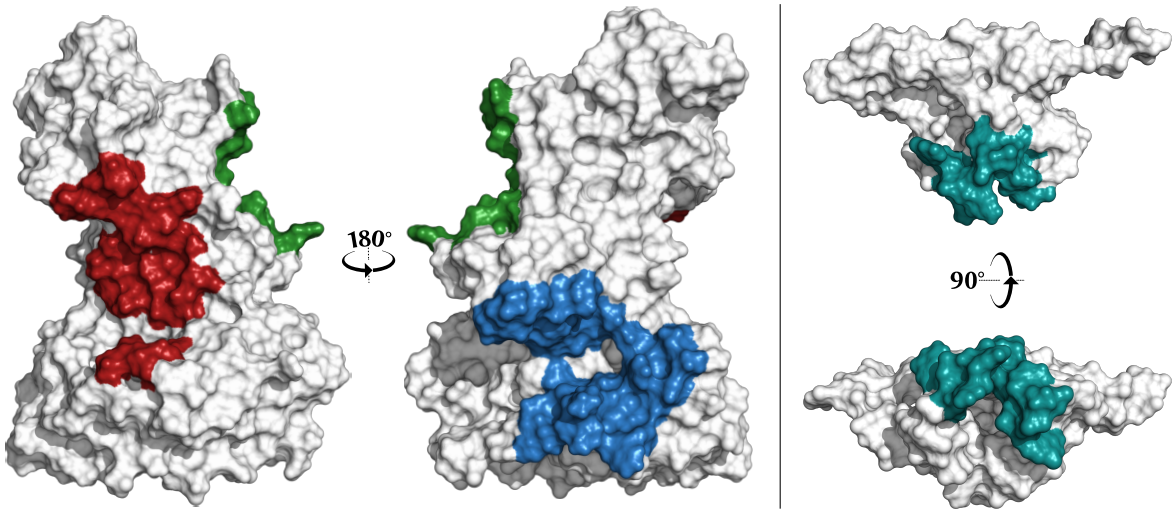


Abb. 6.26: Die Visualisierung der Schnittstellenpräferenzen beider Komplexpartner zeigt die Kernbereiche der bevorzugten Oberflächenregionen in farblicher Kennzeichnung auf der Oberfläche der Moleküle. Auf der linken Seite ist dazu das Zielprotein in der Ansicht aus Abbildung 6.22 sowie von der dazugehörigen Rückseite gezeigt. Die Binderegionen sind auf dem ansonsten weißen Grund grün (Epitop 1), blau (Epitop 2) und rot (Epitop 3) eingefärbt. Auf der rechten Seite ist die kurze Variante des Aptamers sowohl in der Ansicht aus Abbildung 6.23b als auch von der dazugehörigen Unterseite gezeigt. Das Bindemotiv ist hier cyan eingefärbt.

molekularen Editor PyMOL [391]. Die ermittelten passiven Residuen können im Detail in Tabelle 6.11 eingesehen werden. Die AIR selbst liegen in einem speziellen Dateiformat vor, welches paarweise Residuenkontakte jeweils in Kombination mit einem Distanzparameter enthält.

Auch wenn diese Vorgaben damit theoretisch große Freiheiten in der Spezifikation zulassen, wird in der Regel ein festes Format genutzt. Dieses gibt vor, dass alle Verbindungen aus den aktiven Residuen des einen Bindepartners mit den aktiven und passiven des anderen Bindepartners und umgekehrt als interagierende Paare infrage kommen. Für all diese Interaktionen wird ferner der gleiche Distanzschwellwert angenommen. Aus dem Aufbau des Datenformats und der Funktionsweise seiner Anwendung ging hervor, dass über AIR jeweils nur genau eine Bindekonfiguration bestehend aus gegenüberliegenden Einzelflächen beschrieben werden kann. Da auf der Seite des Zielproteins jedoch drei Epitope als mögliche Bindestellen bestimmt wurden, erfolgte die Erzeugung von drei unterschiedlichen AIR, welche jeweils die Verbindung eines Epitops der Proteinstruktur mit dem Sequenzmotiv des Aptamers umfassten.

Informationsgetriebene Dockingsimulation auf Basis der AIR Zwar waren aus der gemeinsamen Betrachtung der primerlosen und -behafteten Variante des Aptamers einige wichtige Hinweise hervorgegangen, mit Beginn der Dockingsimulation erwies sich die vergleichende Betrachtung der langen Aptamervariante jedoch nicht mehr als gewinnbringend. Da sich ferner keine Bindepräferenzen zeigten, wurde die parallele Analyse der langen Aptamervariante zugunsten einer klaren Hauptlinie eingestellt. Basierend auf den Definitionen der drei molekularen Schnittstellen konnten nun informationsgetriebene Dockingsimulationen durchgeführt werden, deren Ergebnisse den Bindungsraum der Komplexe in den bestimmten Ergebnisbereichen höher auflösten und damit zur Identifikation der tatsächlichen Bindekonstellation beitrugen. Die Simulationen umfassten für jede der drei Schnittstellen 1250 entsprechend der Vorgaben eingeschränkt randomisierte, initiale Komplexvarianten. Im Fortgang des Verfahrens wurden jeweils die besten 750 Varianten der semi-flexiblen und die daraus besten 500 der solvatisierten Verfeinerung unterzogen. Eine manuelle Prüfung der erhaltenen Komplexe zeigte, dass die Vorgaben

Tab. 6.11: Die Übersicht der Schnittstellenpräferenzen zeigt für beide Komplexpartner jeweils die Kern- (A) und Randbereiche (P) der Binderegionen, beschrieben mit den zugehörigen Residuenbezeichner und -nummern. Für die kurze Variante des Aptamers war die Region bereits aus den vorhergegangenen Analysen bekannt, während die drei Epitope des Zielproteins aus den am besten bewerteten Komplexen abgeleitet wurden. Die Informationen eignen sich zur Definition der AIR. Eine Visualisierung der Kernbereiche befindet sich in Abbildung 6.26.

Region		Residuen
Apt.	A	T_{18}, G_{19}, G_{20}, T_{21}, C_{22}, C_{23}, G_{24}, G_{25}
	P	G_{13}, A_{14}, G_{15}, G_{16}, A_{17}, G_{26}, G_{27}, C_{28}, G_{34}, A_{35}, G_{36}
Prot.	1A	LYS_{248}, PHE_{250}, THR_{395}, HIS_{396}, GLN_{397}, ARG_{435}, TYR_{444}, ASN_{446}, ASN_{448}, GLN_{504}, HIS_{505}, ASP_{506}, ARG_{536}, ARG_{537}
	1P	GLU_{247}, PRO_{253}, SER_{254}, PHE_{257}, GLN_{390}, GLY_{392}, GLY_{394}, ASN_{398}, PRO_{400}, GLN_{401}, GLN_{402}, TRP_{403}, GLN_{430}, PHE_{433}, THR_{437}, GLY_{443}, PRO_{445}, MET_{447}, ASP_{450}, HIS_{501}, THR_{502}, GLY_{503}, VAL_{508}, ILE_{509}, ASN_{532}, GLY_{533}, THR_{534}, GLY_{535}, ARG_{538}
	2A	MET_{1}, LYS_{2}, MET_{3}, ALA_{4}, SER_{5}, ASN_{6}, ASP_{7}, ASP_{49}, TRP_{51}, ILE_{52}, GLN_{58}, PRO_{87}, TYR_{88}, HIS_{91}, LEU_{92}, ARG_{94}, MET_{95}, PRO_{218}, THR_{219}, PRO_{226}, PHE_{227}, THR_{228}, VAL_{229}, ILE_{231}, ARG_{476}, VAL_{478}, ASN_{479}, PRO_{480}, ASP_{481}, THR_{482}, GLY_{483}, ARG_{484}, VAL_{485}, TYR_{514}, ARG_{516}, PHE_{517}, ASP_{518}
	2P	ALA_{8}, ASN_{9}, ASN_{46}, VAL_{47}, ILE_{48}, PRO_{50}, ARG_{53}, ASN_{54}, ASN_{55}, PHE_{56}, ALA_{59}, PRO_{60}, GLU_{63}, SER_{90}, TYR_{96}, GLN_{106}, ILE_{108}, PRO_{153}, ARG_{201}, LEU_{203}, PHE_{212}, LEU_{215}, VAL_{216}, PRO_{217}, VAL_{220}, GLU_{221}, SER_{222}, LYS_{225}, PRO_{230}, LEU_{232}, THR_{233}, GLU_{236}, HIS_{460}, GLU_{464}, LEU_{486}, PHE_{487}, GLU_{488}, PRO_{510}, PRO_{511}, ASN_{512}, SER_{519}, TYR_{525}, ALA_{539}, LEU_{541}, GLU_{542}
	3A	ILE_{75}, PRO_{129}, THR_{130}, GLU_{131}, GLY_{132}, LEU_{133}, VAL_{258}, GLN_{260}, THR_{267}, ASP_{269}, VAL_{271}, LEU_{273}, GLU_{316}, HIS_{414}, ASN_{415}, VAL_{416}, HIS_{417}, LEU_{418}, ALA_{419}, PRO_{420}, ALA_{421}, VAL_{422}, ALA_{423}, THR_{425}, ASP_{471}, HIS_{492}, LYS_{493}, SER_{494}, TYR_{496}, GLN_{523}
	3P	GLY_{73}, GLU_{74}, TRP_{77}, SER_{78}, PRO_{80}, ASN_{127}, PHE_{128}, SER_{134}, PRO_{135}, GLN_{137}, PHE_{141}, LYS_{177}, ILE_{179}, MET_{181}, PRO_{253}, SER_{254}, GLY_{255}, ALA_{256}, PHE_{257}, GLY_{270}, LEU_{272}, GLY_{274}, THR_{276}, PRO_{313}, THR_{314}, GLU_{315}, ILE_{317}, LEU_{321}, LYS_{361}, TRP_{403}, VAL_{404}, LEU_{405}, PRO_{406}, SER_{407}, SER_{409}, THR_{412}, GLY_{413}, PRO_{424}, PHE_{426}, PRO_{427}, GLY_{428}, GLU_{429}, GLN_{430}, TYR_{462}, ALA_{465}, ALA_{466}, PRO_{467}, ALA_{468}, GLN_{469}, SER_{470}, ASN_{522}, PHE_{524}

für die Interaktionen in allen Fällen eingehalten wurden. Sowohl die großzügige Distanzregelung, die Bestimmung der aktiven Residuen unter Zuhilfenahme mehrerer Kandidatenkomplexe als auch die Zulassung der passiven Residuen in den AIR wirkten in der Simulation als effektive Freiheitsgrade. In den Ergebniskomplexen schlug sich dies in der großen Variabilität der ausgebildeten Schnittstellen nieder, wobei die translatorischen Effekte weniger stark ausgeprägt waren als die der Rotation. Die aufgrund ihrer konkaven Grundwölbung eher als Bindetaschen wirkenden Epitope der Proteinoberfläche ließen in der Bindung prinzipiell wenig Freiheiten für Variationen, was jedoch durch die exponierte Lage der Bindestelle des Aptamers mit seiner hohen Kompaktheit und guten Erreichbarkeit ausgeglichen wurde.

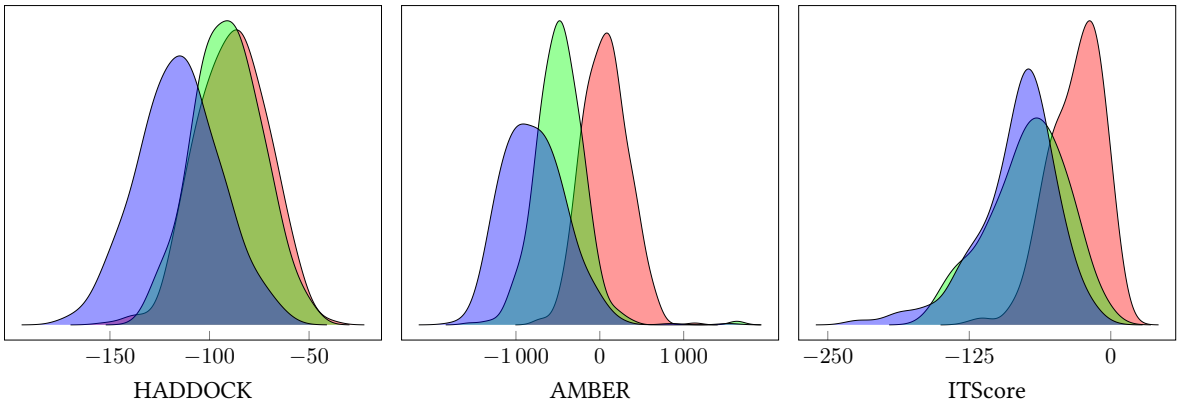


Abb. 6.27: Nach der Durchführung der informationsgetriebenen Dockingsimulation wurden für alle erhaltenen Komplexstrukturen die drei Kennwerte HADDOCK (links), AMBER (mittig) und ITScore-PR (rechts) bestimmt, nach Kennwert getrennt in Form von kernelbasierten Dichteverteilungen aufgetragen und jeweils auf die volle Abbildungshöhe skaliert. In den Randbereichen laufen die Dichteverteilungen systembedingt etwas weiter aus als die Extrema der Wertereihen. Um die drei Epitope der Proteinoberfläche hinsichtlich der ausgebildeten Bindung vergleichen zu können, wurden sie entsprechend dem vorgegebenen Schema aus Abbildung 6.26 eingefärbt. Es zeigt sich generell einheitlich, jedoch in unterschiedlicher Deutlichkeit, dass die Komplexe mit Epitop 2 (blau) vor denen mit Epitopen 1 und 3 (grün und rot) die besten Wertungen erhielten.

Die Güte der drei Proteinepitope wurde nach den erfolgreichen Dockingsimulationen mithilfe der Bewertungsmaße HADDOCK, AMBER und ITScore-PR beurteilt, wobei jedoch der Fokus aufgrund der Erkenntnisse von Kapitel 5 auf dem Letzteren lag. Nach der Bewertung der insgesamt 1500 erzeugten *Decoy*-Strukturen wurden die Werteverteilungen getrennt nach Bewertungsmaß und Epitop aufgestellt. Die zu den drei Epitopen zugehörigen Verteilungen wurden schließlich für jedes Maß zum Vergleich gegenübergestellt. Aufgrund der optisch schlechten Kombinierbarkeit wurde für die Visualisierung in Abbildung 6.27 von Histogrammen abgesehen. Stattdessen wurden auf kernel-basierte Dichteverteilungen zurückgegriffen. Diese laufen in den Randbereichen der Dichteverteilungen systembedingt etwas weiter aus als die Extrema der jeweiligen Wertereihen, erlauben dafür aber eine gemeinsame Darstellung. Wie in der Abbildung erkennbar ist, ließ sich aus allen Bewertungssystemen bezogen auf die Mittelwerte eine einheitliche, wenn auch unterschiedlich stark ausgeprägte Rangfolge der Epitope ableiten. Die Komplexstrukturen mit Beteiligung des Epitops 2 lagen hier vor denen der Epitope 1 und 3. Die Mittelwerte allein konnten hier jedoch kein Entscheidungskriterium sein, da sie nicht repräsentativ für die besten Konstellationen standen und durch die ansonsten sehr großen Überlagerungen der Verteilungen in ihrer Signifikanz beeinträchtigt wurden. Sie konnten jedoch gemeinsam mit der beobachteten Streuung aus statistischer Sicht erste Hinweise über die erhaltenen Schnittstellen liefern. So deutete die klare Trennbarkeit und Ordnung der Verteilungen in der energetischen und sehr von der Größe der Bindeflächen abhängigen Bewertung nach AMBER darauf hin, dass in Verbindung mit Epitop 2 die größten Bindeflächen erreicht wurden. Auch die verlässliche, flächenunabhängige Bewertung nach ITScore-PR widersprach dem nicht, zeigte aber, dass die Epitope 2 und 1 abgesehen von den optimalen Ausläufern in ihrer Verteilung nahezu gleichwertig waren. Neben dieser eher tendenziellen Einschätzung wurden die jeweils am besten bewerteten Komplexe aller drei Proteinepitope in die Betrachtung einbezogen. Wie bereits aus den linken Ausläufern der Verteilungen in Abbildung 6.27 zu erkennen ist, konnten auch hier die Komplexe des Epitops 2 gefolgt von denen der Epitope 1 und 3 die besten Bewertungen vorweisen. Die Güte der drei gefundenen Epitope war daher mit der genannten Einschränkung der Bewertung nach ITScore-PR klar gestaffelt.

6.4.4 Einschätzung der Relevanz für die Epitope der Proteinoberfläche

Das gefundene Aptamer detektiert die Oberflächenproteine des Noroviruskapsids und fällt daher in der Klassifikation der Detektionskonzepte unter die Kategorie der Immunoassays, die aktuell bereits auf Antikörperbasis erfolgreich in der Erkennung des Norovirus eingesetzt werden. In diesem Zusammenhang erlangen neben der theoretischen Bewertung der gefundenen Schnittstellen weitere Relevanzparameter eine hohe Wichtigkeit. Zu diesen gehören die physikalische Zugänglichkeit der Epitope unter realen Umgebungsbedingungen genauso wie eine Einschätzung über die Toleranz der Bindestelle gegenüber Mutationen im Rahmen einer Weiterentwicklung des Virus. An die Simulation schloss sich daher eine genauere Betrachtung der Binderegionen mit Diskussion ihrer praktischen sowie biologischen Relevanz im möglichen Einsatzgebiet der Aptamere an.

Betrachtung der bevorzugten Bindestelle Entsprechend der Bewertungen hatte sich das zweite Epitop als bevorzugte Bindestelle der Proteinoberfläche herausgestellt. Auf der Basis der Bewertung nach ITScore-PR wurde die Struktur mit dem internen Bezeichner `281w` bei einer solchen Wertung von $-229{,}9$ als beste Komplexvariante dieser Bindestelle befunden. Die beiden anderen Bewertungssysteme sprachen dieser Wahl nicht entgegen. So ordnete sich die Bewertung nach HADDOCK mit einem Wert von $-163{,}5$ nur knapp hinter den besten erreichten Ergebnissen dieses Maßes ein. Die rein energetische Bewertung des AMBER-Kraftfeldes blieb zwar mit einem Wert von $-1055{,}6$ etwas weiter hinter den besten Ergebnissen dieses Epitops zurück, befand sich jedoch unabhängig davon sehr klar unterhalb des beobachteten Mittelwertes in einem noch sehr günstigen Bereich. Auch auf struktureller Ebene war die gute Passform des exponierten Aptamer-Bindemusters in der von Epitop 2 beschriebenen Bindetasche des Proteins erkennbar. Die molekulare Schnittstelle ging dabei kaum über die definierten Kernbereiche der beiden beteiligten Epitope hinaus. Wie aus der persönlichen Kommunikation mit dem Leiter der Arbeitsgruppe Norovirus Research der Universität Heidelberg Dr. Grant Hansmann hervorging, konnte die dortige Arbeitsgruppe in Experimenten mit Kapsidproteinen verwandter Norovirusstämme und anderen bindenden Partikeln ebenfalls eine Interaktion im Bereich des Epitops 2 feststellen. Sowohl die theoretische Bewertung, die tatsächliche Bindungsgeometrie als auch die Ergebnisse der Forschergruppe aus Heidelberg wiesen folglich darauf hin, dass es sich bei Epitop 2 um die bevorzugte Bindestelle des Aptamers am `VP1`-Kapsidprotein handelte.

Eine einfache Übertragung dieser Bindepräferenz auf das vollständig assemblierte Viruskapsid war jedoch aus sterischen Gründen nicht möglich. Mit seiner seitlichen, an den Domänenübergang angrenzenden Positionierung auf der `S`-Domäne lag das zweite Epitop vollständig in der Bindetasche, welche die `S`-Domänen aller quasi-äquivalenten Untereinheiten entsprechend der `T=3`-symmetrischen Ikosaederstruktur zur geschlossenen inneren Hülle des Kapsids verbindet. In diesem Kontext war anzunehmen, dass die gute Eignung des zweiten Epitops zur Bindung des Aptamers in Zusammenhang mit der ursprünglichen Aufgabe der Bindetasche in der Polymerbindung des Viruskapsids stand. Eine bereits evolutionär optimierte Geometrie mit günstig angeordneten funktionellen Gruppen könnte die Bindung des Aptamers in diesem Bereich unterstützt haben. Die exklusive Bindung des Aptamers an Epitop 2 hätte jedoch für seine spätere Anwendung in der Detektion des Norovirus eine große Einschränkung bedeutet. So wäre das Aptamer aufgrund der stark begrenzten Verfügbarkeit der Bindestelle ausschließlich in der Lage gewesen, unvollständige Virushüllen mit großer Abhängigkeit vom Grad und der Art ihrer Degeneration zu detektieren. Umso wichtiger wurde es für den späteren Einsatz des Aptamers, eine verlässliche Aussage über die Nutzbarkeit der anderen gefundenen Epitope zu treffen.

Nutzbarkeit der alternativen Epitope Um die Relevanz der anderen Epitope zu überprüfen, wurde durch die Mitarbeiter der TU Dresden eine weitere Aptamerselektion mit dem Verfahren SELEX durchgeführt, wobei jedoch als Zielprotein ausschließlich die P-Domäne des VP1-Kapsidproteins zum Einsatz kam. Auf diese Weise konnte sichergestellt werden, dass lediglich die beiden alternativen Epitope 1 und 3, nicht jedoch das bevorzugte Epitop 2 für die Bindung mit dem Aptamer bereitstanden. Die Anreicherung in der Bibliothek gab daher nicht nur Aufschluss über die Reproduzierbarkeit der Ergebnisse, sondern auch über die Bindefähigkeit des Aptamers an den alternativen Bindestellen. Mit einem pH-Wert von 7,6 wurden analog acht Positiv- und eine finale Negativselektionsrunde durchgeführt, da sich diese Konfiguration in der laufenden Untersuchung durch numerische Stagnation der Anreicherung bereits als ausreichend gezeigt hatte. Die Bibliotheken aller Runden wurden mittels NGS sequenziert und entsprechend den bereits etablierten Qualitätskriterien gefiltert, sodass eine Datenbasis mit vergleichbarer Größe zum Hauptlauf entstand. Die Untersuchung zeigte sowohl in einem Abfall der numerischen Diversitätsbewertungen als auch in der Clusteranalyse eine erfolgreiche Anreicherung von Aptamerkandidaten, die jedoch etwas schwächer ausgeprägt war als im Hauptlauf mit dem gesamten VP1-Protein. Als Gegenprobe wurde die Selektion zusätzlich mit dem Milcheiweiß Kasein durchgeführt, wobei es wie erwartet zu keiner detektierbaren Anreicherung innerhalb der Bibliothek kam.

Um die praktische Relevanz der beiden alternativen Epitope zu prüfen, war eine Anwendung der bioinformatischen Analysewerkzeuge auf den neuen Datensatz nicht zwingend notwendig. Im Zuge einer manuellen Einsicht in die vorsortierten Daten des Clusterings konnte bestätigt werden, dass das Aptamer mit der Insert-Sequenz aus Tabelle 6.6 auch beim alleinigen Einsatz der P-Domäne die höchste Anreicherung in der Bibliothek erreichen konnte. Der exakt identische Fund war dabei auf die gleiche Zusammensetzung der eingesetzten Bibliothek zurückzuführen, die diese spezielle Sequenz in beiden Fällen einschloss. Neben der durchgeführten Negativselektion konnte über die separate Gegenprobe mit Kasein sichergestellt werden, dass es sich bei dem Aptamerfund nicht um einen experimentellen Hintergrundbinder gehandelt hat. Ähnlich wie in der Selektion mit dem gesamten VP1-Protein wurde in nahezu allen höher angereichert vorliegenden Aptamerkandidaten der finalen Bibliothek zudem das Vorkommen des bereits bekannten Motivs TGGTCCGG festgestellt. Dies bestätigte nicht nur die Relevanz des Aptamer- und Musterfundes, sondern in gleicher Weise auch die Limitierung der genutzten Oligonukleotidbibliothek. Bei vollständiger Abdeckung des möglichen Sequenz- und Strukturraumes hätte diese mit sehr hoher Wahrscheinlichkeit andere Aptamerkandidaten mit einer besseren Eignung für die veränderte Oberflächensituation des Zielmoleküls enthalten. Die Untersuchung belegte jedoch in erster Linie, dass das Epitop 2 nicht die alleinige Bindestelle des Aptamers am VP1-Kapsidprotein sein konnte, sondern auch im Bereich der P-Domäne eine Bindung mit dem Aptamer möglich war. Die geringere Anreicherung der untersuchten Aptamersequenz bei Verwendung der P-Domäne als Zielprotein stand dabei in Einklang mit der geringeren Wertung der alternativen Epitope 1 und 3 in den Simulationsmodellen mit dem gesamten VP1-Protein.

Betrachtung der alternativen Epitope Unter den verbliebenen zeigte das Epitop 1 nahezu in allen untersuchten Bewertungskriterien die besseren Ergebnisse, sowohl in Mittelwert und Form der erreichten Verteilungen als auch in den jeweiligen Extrema. Allein in der Bewertung nach HADDOCK zeigte sich nur ein knapper Vorsprung für Epitop 1 in der Gesamtverteilung sowie eine einzelne bessere Komplexstruktur im konkurrierenden Epitop. Da sich das erste Epitop im oberen Bereich der P1-Domäne an der Rückseite zur beginnenden P2-Domäne befand, genoss

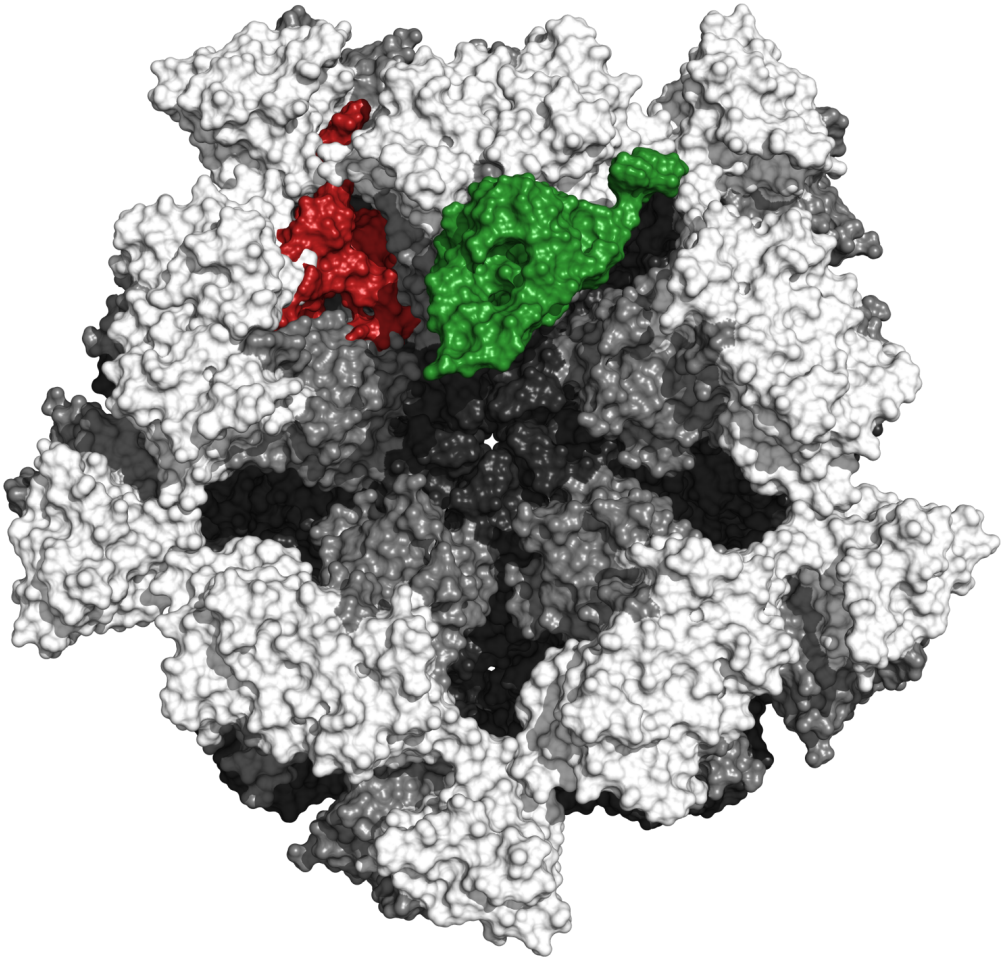


Abb. 6.28: Während Epitop 2 im assemblierten Noroviruskapsid vollständig blockiert ist (nicht sichtbar), unterscheiden sich die Zugänglichkeiten der beiden verbliebenen Epitope für das Aptamer markant. Um diese Unterschiede darzustellen, wurde eine der 32 kelchförmigen Vertiefungen aus der Gesamtansicht des Viruskapsids aus Abbildung 6.3b herausgegriffen und zur Einbettung des Aptamers in seinen zwei Bindepositionen genutzt. Zur klareren Differenzierung zwischen Protein und Aptamer wurde das Farbschema der Proteindomänen auf Helligkeitsabstufungen verringert (S-Domäne in dunkelgrau, P1-Domäne in mittlerem grau, P2-Domäne in hellgrau). Die Positionen des Aptamers an den beiden Epitopen 1 (grün) und 3 (rot) wurden durch ein Alignment zwischen den jeweils simulierten Komplexen mit einer der VP1-Einheiten der dargestellten PDB-Struktur 1IHM [458] bestimmt. In der Abbildung wird deutlich, dass für das Aptamer an Epitop 1 (grün) keine Einschränkungen bei der Zugänglichkeit auftreten. Dies gilt auch bei der Bindung mehrerer Aptamere innerhalb einer Vertiefung (maximal fünf mit leichter struktureller Anpassung). An Epitop 3 (rot) kommt es hingegen sowohl beim Zugang als auch bei der finalen Positionierung des Aptamers in der Bindestelle zu bedeutenden sterischen Kollisionen, die einen energetischen Nachteil für dieses Epitop bedeuten.

es auch im assemblierten Viruskapsid eine hervorragende Zugänglichkeit für das bindende Aptamer. Zwischen den anderen am Aufbau des Kapsids beteiligten Kopien des VP1-Proteins und dem Aptamer waren dabei keine sterischen Kollisionen zu beobachten, was sich bei der Bildung des Komplexes energetisch sehr günstig auswirkt. Abbildung 6.28 gibt einen Eindruck über die Lage des Aptamers auf einem größeren Ausschnitt der Kapsidoberfläche. Epitop 1 eignet sich daher besonders gut für die Detektion des Norovirus sowohl in intakter als auch in degradierter Form. In einer der 32 Vertiefungen der Virusoberfläche können bis zu fünf Aptamere binden, wobei jedoch leichte sterische Überschneidungen bei voller Besetzung zu einer minimalen strukturellen Anpassung der Bindekonstellation anregen. Die räumlichen Gegebenheiten erlauben dabei theoretisch sogar eine Erweiterung des Aptamers um eine kovalent gebundene Nutzlast,

mit deren Hilfe der Transport eines Wirkstoffes oder einer speziellen Markierung zum Virus durchgeführt werden könnte. Da Epitop 1 nicht bis in die hochvariable Region der P2-Domäne hineinreicht, kann darüber hinaus angenommen werden, dass das Aptamer mit seiner Bindung an diesem Epitop auch in der Lage ist, nahe verwandte Norovirusstämme zu erkennen.

Epitop 3 nahm in den angestellten Bewertungen fast ausnahmslos den letzten Rang ein. Durch seine Positionierung im Übergangsbereich von der S- zur P1-Domäne war es zwar nicht durch die Polymerbildung der VP1-Proteine blockiert, allerdings sehr tief in der Oberflächenstruktur eingelassen. Innerhalb der typischen, kelchförmigen Senken war in dieser Tiefe jedoch nicht ausreichend freier Raum für die tatsächliche Anlagerung des Aptamers verfügbar. Von den sechs angrenzenden Proteineinheiten waren zwei ohne Berührung und zwei weitere mit guter Passform der Berührungsfläche weitestgehend unproblematisch für die Bindung des Aptamers. Zwei weitere Proteineinheiten bildeten in Verbindung mit dem eingesetzten Aptamer deutliche sterischen Kollisionen aus. In der Folge wäre daher sowohl zum Erreichen der Bindestelle als auch zur Bindung des Aptamers an Epitop 3 eine strukturelle Anpassung dieser beiden Proteineinheiten und des Aptamers unter Aufwendung von Energie notwendig, was gegen die praktische Relevanz des dritten Epitops in der Norovirusdetektion sprach. Zur visuellen Unterstützung zeigt Abbildung 6.28 die Lage des Aptamers im Kontext der Kapsidoberfläche. Die Interaktionen mit den umliegenden Proteineinheiten blieben sowohl in der experimentellen Anordnung als auch in den bioinformatischen Analysen unbeachtet, sodass eine verlässliche Aussage an dieser Stelle schwierig zu treffen war. In Hinsicht auf die günstige Lage und Bewertung des ersten Epitops wurde jedoch die Vermutung gestützt, dass es im assemblierten Viruskapsid ausschließlich zu einer Bindung des Aptamers an Epitop 1 kommt. An den Bruchstellen eines degradierten Kapsids besteht jedoch weiterhin die Möglichkeit einer Bindung mit Epitop 3.

Schlussbetrachtung Zusammenfassend kann bemerkt werden, dass es auf Basis der erhobenen Sequenzdaten mithilfe komplexer Analyse-, Vorhersage- und Simulationsstrategien möglich war, die Bindung zwischen Aptamer und VP1-Kapsidprotein des Norovirus näher zu charakterisieren. Die Oberfläche des Proteins konnte dabei auf drei spezifische Epitope verringert werden, von denen genau eins, Epitop 1, für die Bindung am assemblierten Kapsid infrage kam. Am freigestellten VP1-Protein sowie an den Bruchkanten degradierter Kapside ist darüber hinaus auch die Bindung über die anderen beiden Epitope wahrscheinlich. Bereits in der Analyse der Sequenzdaten der Aptamerbibliothek stellte sich das Motiv TGGTCCGG durch statistische Häufung deutlich heraus, welches sich auch in den Ergebnissen der Sekundär- und Tertiärstrukturvorhersage sinngerecht passend niederschlug. Dieses Motiv konnte durch die weiteren Analysen dieses Unterkapitels als bindungsrelevant eingestuft werden. Durch die einzelne Betrachtung der infragekommenden Epitope im Kontext ihres späteren Einsatzes konnte ferner die experimentell bestätigte Bindefähigkeit des Aptamers für das einzelne VP1-Protein auf das vollständig assemblierte Viruskapsid übertragen werden.

Rückblickend können jedoch auch einige Verbesserungen am experimentellen Vorgehen vorgeschlagen werden. Diese betreffen besonders die Wahl des Zielobjekts für die Selektion. Wie sich zeigte, bietet der Grundbaustein VP1 des Noroviruskapsids mehrere Bindeflächen, die im assemblierten Kapsid nicht mehr für das Aptamer zugänglich sind. Bei der Auswahl und Vorbereitung des Zielmoleküls sollte diese Tatsache Beachtung finden, um die Selektion von praktisch nicht nutzbaren Aptameren von vornherein zu vermeiden. Im vorliegenden Fall wäre ein Teil der nicht-relevanten Bereiche durch die Verringerung des Zielproteins auf die P-Domäne weggefallen. Aber auch in diesem Fall verbliebe die nicht zugängliche Unterseite dieser Domäne, die nur

durch eine mechanische Blockierung, im Rahmen der Immobilisierung oder durch zusätzliche Negativselektion als Zielepitop vollständig ausgeschlossen werden kann. In der bioinformatischen Analyse zeigte sich ferner, dass die Bewertung eines Epitops ohne Berücksichtigung des weitreichenden molekularen Kontextes nicht zwangsläufig auf dessen praktische Nutzbarkeit übertragbar ist, da in der Selektion mit Einzeleinheiten einige wichtige Effekte des Gesamtkomplexes nicht auftreten. Neben sterischen Behinderungen mit anderen VP1-Einheiten waren dies primär konkurrierende und ergänzende Interaktionen im komplexen Umfeld. Um während der Selektion der Aptamere auf diese geometrischen und physikochemischen Gegebenheiten des Gesamtkomplexes eingehen zu können, ist die Nachbildung der tatsächlichen Makrostruktur zumindest in relevanten Anteilen unerlässlich. Es ist dafür nicht unbedingt notwendig, vollständig zusammengesetzte Viruskapside als Zielobjekte in die Selektion einzubringen, da bereits geeignet gewählte Ausschnitte der Oberfläche für eine Anlagerung der Aptamere unter realitätsnahen Bedingungen ausreichen. Auch hier ist jedoch zu beachten, dass praktisch unzugängliche Flächen von der Selektion entsprechend ausgeschlossen werden sollten.

Als weiterer Angriffspunkt für Optimierungen zeigte sich die eingesetzte Oligonukleotidbibliothek, deren Limitierung in der Untersuchung deutlich wurde. Obwohl Epitop 2 in der Kontrolluntersuchung ausgeschlossen wurde, befand sich innerhalb der Bibliothek kein Aptamer, welches eine höhere Affinität zu den verbliebenen Epitopen 1 und 3 aufwies. Da der Versuch, größere Oligonukleotidbibliotheken einzusetzen, noch lange vor Erreichen einer annähernd vollständigen Abdeckung des Sequenz- und Strukturraumes an die experimentellen Grenzen stößt, muss die Optimierung hier andere Wege einschlagen. Die geradlinige Überlegung, über die Kombination von unterschiedlichen Bibliotheken in aufeinanderfolgenden Selektionsexperimenten möglichst überschnittsfreie Teilräume des Sequenz- und Strukturraumes zu vereinen, scheitert praktisch an dem hohen experimentellen Aufwand. Ebenfalls ein Eingriff in das experimentelle Geschehen und zugleich ein Paradigmenbruch ist die Einführung gezielter Mutations-Phasen. Diese erhöhen ähnlich eines tatsächlichen evolutionären Prozesses die Diversität der Bibliothek gezielt in denjenigen Teilbereichen des Sequenz- und Strukturraumes, die sich in der bereits erfolgten Selektion als relevant herausgestellt hatten. Im vorliegenden Fall bot sich darüber hinaus an, die rein zufällig generierte Bibliothek durch eine targetspezifisch angereicherte und damit semi-randomisierte Bibliothek zu ersetzen. Auch diese erreicht trotz geringer Gesamtabdeckung eine höhere strukturelle Auflösung im bereits als relevant bestimmten Unterraum des Sequenz- und Strukturraumes. Die Basis für diese Anreicherung wurde bereits durch die Aptamerselektion geschaffen.

6.5 Konzepte für die spezifische Anreicherung von Oligonukleotidbibliotheken

Im Verlaufe einer Aptamerselektion führt die sukzessive Verringerung des Sequenz- und Strukturraumes der eingesetzten Bibliothek zu einer Veränderung ihres informationellen Gehaltes. Nach Abschluss der Selektion enthält die verwendete Bibliothek schließlich wichtige Informationen zu den Charakteristika der bindenden Aptamere. Diese stehen über die kausalen Zusammenhänge der biologisch-chemischen Erkennung mit entsprechenden Eigenschaften des Zielproteins in Verbindung und zeichnen sich damit durch eine hohe Relevanz aus. Da diese Informationen jedoch in der Komplexität der Bibliothek verborgen liegen, sind sie auf direkte Art nur sehr eingeschränkt nutzbar. Mithilfe der bioinformatischen Analysemethoden ist es jedoch möglich, Kerninformationen aus der Bibliothek zu extrahieren und damit für deren Optimierung

verfügbar zu machen. Die Berücksichtigung von Aspekten der Validierung während der Analyse wirkt sich dabei sowohl auf die Zuverlässigkeit der extrahierten Informationen als auch auf den Erfolgsgrad der Bibliotheksoptimierung positiv aus. Der Erfolg wird hierbei über das Erreichen einer hohen Auflösung des Sequenz- und Strukturraumes in denjenigen Bereichen definiert, die potente Aptamerkandidaten für das gewählte Zielmolekül enthalten. Dies kann primär über die Erhöhung der Auflösung in bereits bekannten sowie über die Erschließung bisher unbeachtet gebliebener, neuer Sequenz- und Strukturbereiche geschehen.

6.5.1 Ableitbare Unterräume des Sequenz- und Strukturraumes

Werden die Sequenzdaten der finalen Bibliothek sowie die Ergebnisse der bioinformatischen Analyse zugrunde gelegt, so ergeben sich unterschiedliche Strategien für die Ableitung derartiger Unterräume. Ihre Nutzung richtet sich prinzipiell nach dem konkret vorliegenden informationellen Bedarf und den wissenschaftlichen Zielvorstellungen. Die Bedeutung, die Bereichen innerhalb als auch außerhalb von Motivfunden im Kontext der Analyseergebnisse zugemessen wird, wirkt als kritischer Einflussfaktor bei der Wahl und Parametrisierung dieser Strategien. In den folgenden Absätzen wird eine Auswahl der benannten Strategien auf Basis einer einfachen Systematisierung vorgestellt.

Ausschließliche Nutzung der Sequenzdaten der finalen Bibliothek Da die Anreicherung der Aptamere mit dem Ende des Selektionsexperimentes soweit abgeschlossen war, kann durch die unverarbeitete Weiternutzung der sequenzierten Bibliothek keine Optimierung erreicht werden. Die bestehenden Sequenzen können jedoch durch zufällige Mutationen zu einer semi-randomisierten Bibliothek entwickelt werden, wobei unterschiedliche Mutationsstrategien möglich sind. Über feste Mutationsquoten kann beispielsweise die Anzahl der Mutationen pro Sequenz vorgegeben werden, sodass nur deren Arten und Positionen zufällig bestimmt werden. Ein natürlicheres Verhalten wird durch den Einsatz fester Mutationsraten erreicht, indem alle Nukleobasen unabhängig voneinander behandelt werden und damit auch die Anzahl der Mutationen pro Sequenz dem Zufall unterliegt. Zwar ist auch eine kontextsensitive, variable Vergabe der jeweiligen Mutationsquoten und -raten denkbar, jedoch bieten die unverarbeiteten Sequenzdaten über die einfach bestimmbaren Häufigkeiten hinaus kaum weitere Ansatzpunkte für eine qualifiziert Parametrisierung. Da nur ein Bruchteil der finalen Bibliothek von der Sequenzierung erfasst wurde, kann von jeder ihrer Sequenzen eine Vielzahl von Mutanten generiert werden, ohne dabei die technische Größengrenze zu überschreiten. Auch wenn dadurch eine gute Streuung um die bisherig für gut befundenen Strukturen erreicht werden kann, muss bemerkt werden, dass ein solches Vorgehen kaum zielgerichtet ist. Mit einer Reduktion auf einen kleinen Teil der häufigsten Sequenzen der finalen Bibliothek kann zwar eine detailliertere Ausleuchtung des Sequenz- und Strukturraumes um tendenziell bedeutungsvollere Aptamerkandidaten erreicht werden, aber auch hier wird lediglich eine leichte Refokussierung auf relevante Eigenschaften der Sequenzen erreicht. Insgesamt ist das Verfahren damit zwar nicht anfällig für analysebedingte Fehler, es kann jedoch auch keinen Gewinn aus ihren Ergebnissen ziehen.

Integration bioinformatischer Analyseergebnisse in die Mutationsstrategie Ein Ausweg besteht in der Erweiterung der eben vorgestellten Verfahren um Informationen aus der bioinformatischen Analyse. Diese erlauben über die Parametrisierung variabler Mutationsraten die kontextsensitive Steuerung der sequenziellen und damit strukturellen Streuung sowohl pro Sequenz als auch pro Nukleobase. Eine kritische Rolle spielt dabei die Aufteilung der Sequenz

in Regionen unterschiedlicher Bestimmung. Relativ offensichtlich sind hierbei die Bindemotive, deren physikochemische Eigenschaften und intrinsische Strukturierung sich für die Bindung auf einem speziellen Epitop des Zielproteins besonders gut eignen. Während die Bedeutung im Falle eines einzelnen Bindemotivs relativ klar ist, muss die Analyse im Falle mehrerer Bindemotive ebenfalls Auskunft darüber geben, inwiefern diese kausal miteinander in Verbindung stehen. Die gemeinsame Adressierung eines einzelnen Epitops ist dabei beispielsweise anders zu behandeln als das Vorhandensein alternativer Kombinationen aus Bindemotiven und Epitopen. Am konkreten Aptamer-Target-Komplex ist jedoch nicht allein das Bindemotiv beteiligt, da dieses nur dann mit dem zugehörigen Epitop interagieren kann, wenn die Gesamtstruktur des Aptamers eine entsprechende Positionierung ermöglicht. Die Bereiche außerhalb der detektierten Motive können aus diesem Grund selbst dann Einfluss auf die Bindefähigkeit und Affinität eines Aptamers haben, wenn sie selbst nicht explizit in der Bibliothek angereichert vorliegen.

Über eine gezielte Einflussnahme auf die Mutationsraten ist es auf dieser Basis möglich, den Sequenz- und Strukturraum sowohl im Bereich der Binderegionen als auch in dem der Gesamtstruktur zu erweitern. Praktisch lässt sich das bereits durch fallspezifisch leicht unterschiedlich gewählte Mutationsraten in den jeweiligen Bereichen umsetzen. Stärkere Unterschiede in den Raten können dabei bis zum vollständigen Erhalt sowie der vollständigen Randomisierung einiger Teilbereiche führen, um zunehmend schärfere Abgrenzungen zu erreichen. Die großen Freiheiten dieses Verfahrens erfordern eine gute Konzeption auf Basis der vorherigen Analyseergebnisse, unterliegen jedoch bezüglich der positionellen Einbettung der Muster in der Gesamtsequenz trotz allem den Vorgaben der sequenzierten Bibliothek.

Generierung einer strukturell unabhängigen Bibliothek Soll auch von diesen Vorgaben abgewichen werden, verbleibt die Möglichkeit, eine Bibliothek allein entsprechend den Vorgaben aus der bioinformatischen Analyse zu erzeugen. Ein Erhalt der Gesamtstruktur ist dabei in der Regel nicht möglich, sodass sich das Augenmerk in diesem Fall auf die Variation und neue Einbettung der Bindemotive richtet. Während die Variation der Motive nach den bekannten Verfahren geschehen kann, sind für die Einbettung in eine neue Gesamtstruktur Templatesequenzen notwendig. Diese setzen sich neben den Motiven aus zufällig generierten Segmenten konstanter und variabler Länge zusammen und erlauben damit eine freie Positionierung der Motive entlang der gesamten Aptamersequenz. Durch die großen Anteile zufälliger Sequenzen ist jedoch eine Interaktion zwischen diesen und den eingeführten Motiven sehr wahrscheinlich, wodurch die intrinsische Struktur der Motive in zahlreichen Fällen beeinträchtigt oder gar zerstört werden kann. Es ist daher in der Planungsphase einer solchen Bibliothek angeraten, die intrinsische Struktur der Bindemotive durch ihre Einbettung in unterstützende Strukturelemente zu stabilisieren. Ein typisches Beispiel hierfür ist die Stabilisierung einer *Hairpin*-Struktur durch einen umschließenden, helikalen Stem-Bereich.

6.5.2 Zusammensetzung einer Bibliothek

In den drei vorgestellten Grundstrategien lassen sich durch unterschiedliche Parametrisierung eine Vielzahl verschiedenartig spezialisierter Unterräume des Sequenz- und Strukturraumes ableiten. Nicht in jedem Fall kann die Analyse jedoch eine eindeutige Präferenz zur Auswahl einer solchen Strategie und Parametrisierung liefern. Es sollte daher in Fällen uneindeutiger oder fehlender Präferenz die Kombination mehrerer dieser Unterräume in Betracht gezogen werden, um ein zufriedenstellendes Ergebnis zu erreichen. Dieses besteht in der theoretischen Beschreibung einer targetspezifisch optimierten Oligonukleotidbibliothek, welche anstatt ihres vollständig

zufällig generierten Pendants in der Aptamerselektion eingesetzt werden kann, um höhere Affinitäten zum Zielprotein zu erreichen. Als letzter Schritt vor ihrem produktiven Einsatz muss die Synthese der optimierten Bibliothek erfolgen. Da der damit verbundene Aufwand prinzipiell mit der Komplexität der theoretischen Zusammensetzung der Bibliothek wächst, sollten die technischen Gegebenheiten des Syntheseprozesses bereits während der theoretischen Konzeption der Bibliothek im Blick behalten werden. Die Verfahren der künstlichen Nukleotidsynthese sind technisch mittlerweile jedoch so fortgeschritten, dass auch die gezielte Mutation und die Anreicherung der Bibliothek mit festen Sequenzen keine Hindernisse mehr darstellen [565; 566]. Sowohl die technischen als auch die analytischen Möglichkeiten sind damit so geschaffen, dass mithilfe ihres Zusammenspiels das Potential der Aptamere und Aptamerselektion tiefer ausgeschöpft werden kann.

7 Abschließende Betrachtung

Nach der umfänglichen Bearbeitung der gesetzten inhaltlichen Schwerpunkte sollen die Ergebnisse dieser Arbeit im Rahmen der abschließenden Betrachtung dieses Kapitels kurz zusammengefasst und perspektivisch eingeordnet werden. Die Zusammenfassung der Ergebnisse orientiert sich dabei an den einführend aufgestellten vier Hypothesen, die als Ankerpunkte im Prozess der Wissensfindung fungierten, und nimmt Bezug auf die untergeordneten Fragestellungen. Die Ausführung endet jeweils mit der Verifikation oder Falsifikation der Hypothese.

7.1 Zusammenfassung der Ergebnisse

Mit ihrer Fähigkeit zur Bindung eines breiten Spektrums möglicher Zielmoleküle besitzen Aptamere ein großes Anwendungspotential in der biotechnologischen Industrie und der Medizin. Die methodischen Grenzen ihres Herstellungsverfahrens SELEX stellen die Anwender jedoch vor das entscheidende Problem, dass die selektierten Aptamere mit großer Wahrscheinlichkeit nur eine suboptimale Affinität und Spezifität zum Zielmolekül aufweisen. Obwohl auch diese suboptimalen Aptamere bereits praktisch eingesetzt werden, ist die Selektion von Aptameren mit höherer Affinität und Spezifität erstrebenswert. Diese leistungsfähigeren Alternativen können in geringeren Konzentrationen, mit weniger ungewollten Seiteneffekten und damit deutlich verlässlicher und effektiver angewendet werden. Ein Weg zur Identifikation verbesserter Aptamerkandidaten führt über die Bindungsinformationen, die sich während eines Selektionsprozesses in der eingesetzten Bibliothek niederschlagen. Die bioinformatische Erschließung dieser Bindungsinformationen wurde daher anhand zweier inhaltlicher Schwerpunkte zur Zielstellung dieser Arbeit. Diese bezogen sich auf die Suche und Evaluation relevanter, bioinformatischer Analysemethoden und die kombinierte Anwendung der Analyseverfahren am konkreten Beispiel.

Hypothese 1 Die physikochemische Prägung numerischer Deskriptoren ist bei der Charakterisierung von Aptamersequenzen entscheidend für die Güte der abgeleiteten Beschreibung.

In der ersten Hypothese wurden die Primär- und Sekundärstruktur als Ebenen der Beschreibung der Aptamere herausgegriffen. In Form von Sequenzdaten und Sekundärstrukturvorhersagen können diese im Rahmen der gegebenen Vorhersagegenauigkeit relativ einfach auch in größerer Zahl bereitgestellt werden. Für die Untersuchung war darüber hinaus ein Referenzwert für die Wirkstärke der Sequenzen essentiell. Da keine hinreichend große Menge von Aptamersequenzen zu einem speziellen Zielprotein gefunden werden konnte, die hinsichtlich ihrer Affinitäten charakterisiert waren, kam ein Datensatz von Promotorsequenzen als alternative Form funktioneller Nukleinsäuren zum Einsatz. Die numerische Beschreibung der Nukleinsäuren erfolgte prinzipiell über zwei Wege. Der indirekte Beschreibungsweg verlief über die initiale Charakterisierung der Nukleobasen und die anschließende Anwendung dieser auf Nukleinsäuren mithilfe spezieller Transformationsverfahren. Sowohl die Nukleobasendeskriptoren als auch die Transformationen wurden der Literatur entnommen. Über molekulare Deskriptoren der ato-

maren Ebene und Worthäufigkeiten, sogenannte n-Gramme, kann jedoch auch eine direkte Beschreibung der Nukleinsäuren ohne intermediäre Nukleobasendeskriptoren erreicht werden. Die Deskriptoren dieser beiden Gruppen stützten sich sowohl auf absolute und relative Positionsinformation als auch auf explizite und implizite physikochemische Information. Die Güte eines Deskriptorensets wurde an seiner Fähigkeit gemessen, die biologische Realität sinnvoll *in silico* abzubilden. Um dies zu überprüfen, wurden auf Basis der numerischen Beschreibung und der gegebenen Wirkstärken der Promotoren Regressionsmodelle erzeugt und miteinander verglichen. Die Ergebnisse erlaubten nicht nur eine Beurteilung der Deskriptorengüte, sondern ließen auch Rückschlüsse auf die Einflussnahme der informationellen Komponenten zu. Die Kombination aus impliziter physikochemischer Information und Positionsinformation führte dabei zu einer Beschreibung von hoher Güte, wie sich an den geringen Vorhersagefehlern der korrespondierenden Regressionsmodelle erkennen ließ. Da die Einteilung in Nukleobasen eine hinreichende Menge impliziter physikochemischer Information bereitstellte, konnte durch die Anreicherung der Deskriptoren mit expliziter physikochemischer Information kein weiterer Gewinn erzielt werden. Die Vernetzung der physikochemischen Informationen wurde durch Positionsinformationen in passender Größenordnung sichergestellt und wirkte sich maßgebend auf die Güte der Beschreibung aus. Das Fehlen dieser Vernetzung manifestierte sich im Deskriptorenset PaDEL in deutlich höheren Modellfehlern und dem Ausbleiben des positiven Effektes bei der *Feature Selection*. Im Rahmen dieser Untersuchung stellten sich die n-Gramm-Deskriptoren, deren Beschreibung lokaler Sequenzfragmente gut mit dem Bindungsprinzip der funktionellen Nukleinsäuren korrespondiert, als ideale Form der Beschreibung heraus. Die Verwendung der Sekundärstrukturen und die Kombination unterschiedlicher Wortlängen waren dafür Bedingung. Da bei der Charakterisierung von Sequenzen funktioneller Nukleinsäuren kein statistisch-evidenter Zusammenhang zwischen der physikochemischen Prägung der verwendeten Deskriptoren und der erreichbaren Modellgüte festgestellt werden konnte, gilt die erste Hypothese als widerlegt.

Hypothese 2 Auch unter der Anforderung von Variabilität ist die Mustersuche in großen Datensätzen von Aptamersequenzen mit begrenzter Rechenkapazität durchführbar.

Die Auswertung der beiden Beschreibungsebenen der Nukleinsäurestruktur wurde in der zweiten Hypothese in Hinblick auf statistisch-symbolische Signale weitergeführt. Zur verlässlichen Auswertung war hier ein größerer Sequenzdatensatz notwendig, dessen Sequenzen jedoch nicht hinsichtlich ihrer Wirkstärken annotiert sein mussten, da kein biologisches Referenzkriterium in die Auswertung einfloss. Für die Mustersuche in großen Sequenzdatensätzen eignete sich als Datenstruktur in besonderem Maße der Suffixbaum. Dieser kann unter Eingabe aller Sequenzen in vertretbarer Zeit erzeugt werden und erlaubt die schnelle Suche nach überrepräsentierten Subsequenzen unabhängig von deren sequenzieller Positionierung. Die Subsequenzen dürfen dabei keine Variabilität enthalten und entsprechen damit der eingeführten Definition trivialer Muster. Auf algorithmischer Seite kann zwar die Geschwindigkeit der Baumerzeugung optimiert werden, praktisch ist dies aber unter den gegebenen Umständen nicht notwendig gewesen. Die Variabilität biologischer Sequenzen wurde durch die Erweiterung der Musterpositionen um die Fähigkeit der mehrdeutigen Zuordnung in die bestehende Definition trivialer Muster eingeführt. Während die gezielte Nutzung derartiger Mehrdeutigkeiten die Aussagekraft der betrachteten Muster verbessert, führt ihr übermäßiger Einsatz zur Beliebigkeit und damit auch zur Bedeutungslosigkeit der Muster. Die verlässliche Quantifizierung und sinnvolle Restriktion der Variabilität während der Mustersuche war daher ein wichtiges Anliegen. Das eingeführte

Konzept der Variabilität konnte jedoch nicht in den Suffixbaum integriert werden, da sich in diesem Fall das exponentielle Speicherverhalten der Datenstruktur in einen nicht handhabbaren Bereich entwickelte. Es wurde daher über eine Erweiterung des bestehenden Suchverfahrens nach dem Prinzip des progressiven *Node Merging* eine algorithmische Lösung entworfen, die zum Erreichen eines akzeptablen zeitlichen Verhaltens moderate bis strenge Begrenzungskriterien erfordert. Da jedoch auch die Aussagekraft der gefundenen Muster mit der Strenge dieser Kriterien korrespondiert, stellen diese praktisch keine Einschränkung für das Verfahren dar. Ein wichtiger limitierender Faktor ist hingegen die Größe des zugrundeliegenden Alphabets, da dieses den Umfang der möglichen Variabilität maßgeblichen bestimmt. Während hier bei Nukleinsäuresequenzen keine Probleme auftreten, müsste für die Mustersuche in Proteinsequenzen beispielsweise eine einschränkende Definition möglicher variabler Musterpositionen zwingend vorgenommen werden, um die Suche durchführen zu können. Da die Mustersuche in großen Datensätzen von Aptamersequenzen unter der Anforderung von Variabilität mit begrenzter Rechenkapazität durchführbar ist, wenn hinreichend strenge Begrenzungskriterien vorgegeben werden, gilt die zweite Hypothese als bestätigt.

Hypothese 3 Paarweise Bewertungsfunktionen können die interatomaren Kontakte zwischen Proteinen und Aptameren selbst dann zur Beurteilung eines simulierten Komplexes nutzen, wenn keine Referenzstruktur zum Vergleich vorliegt.

In der dritten Hypothese wechselte die Betrachtung in die Ebene der Tertiärstruktur, was die Datengrundlage und damit auch die Ausrichtung der Analyse deutlich veränderte. Sowohl bei der experimentellen Aufklärung als auch bei der computergestützten Simulationen von Tertiärstrukturen handelt es sich um hochkomplexe Prozesse, die mit einem erheblichen methodischen Aufwand verbunden sind. Aus diesem Grund bleibt die Verfügbarkeit von Tertiärstrukturen für die Aptamere auf wenige Exemplare beschränkt. In der Ebene der Tertiärstruktur war eine alleinige Untersuchung zahlreicher Aptamerstrukturen infolgedessen kein zweckmäßiges Mittel zur Charakterisierung einer gemeinsamen Binderegion. Vielmehr lag nun die Aufklärung der konkreten Bindegeometrie zwischen einem Aptamer und dem zugehörigen Zielprotein im Mittelpunkt des Interesses. Nach der Simulation der Tertiärstrukturen der beiden Bindepartner erfolgt dazu eine Docking-Simulation, die eine Menge von Kandidatenkomplexen liefert. Dem nahenativen Bereich werden hierbei diejenigen Komplexe zugeordnet, die eine hohe Ähnlichkeit zur in der Regel unbekannten natürlichen Bindegeometrie aufweisen, während der nicht-native Bereich diejenigen Strukturen vereint, die sich in Position oder Orientierung davon unterscheiden. Die Auswahl nahe-nativer Komplexe war damit die Priorität der bioinformatischen Analyse. Eine Gruppe infrage kommender Bewertungsmodelle waren die wissensbasierten Paarpotentiale, die sich auf die statistische Analyse bekannter Protein-Nukleinsäure-Komplexe gründen. Die Aussagekraft der meisten dieser Ansätze litt jedoch stark unter der Reduktion der Modellkomplexität, der Vernachlässigung von Informationsgruppen und der geringen Größe und Repräsentativität der Datengrundlage. Die zwei Ansätze SPA-PN und ITScore-PR setzten sich durch ein Optimierungsverfahren positiv von den anderen Vertretern der ersten Gruppe ab und wurden in der Evaluation daher weiter verfolgt. Die molekularmechanischen Bewertungsmodelle bildeten eine zweite Gruppe, die auf Basis quantenmechanischer Berechnungen und experimenteller Messungen die physikalischen Interaktionen der molekularen Schnittstelle approximieren. Die Auswahl von Vertretern ist in dieser Gruppe aufgrund der hohen Ähnlichkeit von Datenbasis und Grundkonzept eher formal geschehen. Sie umfasste infolge seiner guten Reputation das Kraftfeld AMBER und aus praktischer Erwägung das auf OPLS basierende, interne

Bewertungsmodell der eingesetzten Docking-Software HADDOCK. Für die Evaluation wurden zwei unabhängige Referenzkomplexe aus Protein und Aptamer gewählt, die weder im Trainings- noch im Testdatensatz der vier gewählten Bewertungsmodelle involviert waren und molekulare Schnittstellen unterschiedlicher Größe ausprägten. Nach der Erzeugung von *Decoy*-Komplexen mithilfe von HADDOCK zeigte sich für die strukturelle Abweichung der Komplexe von der Referenz ein RMSD-Wert von $10\,\text{Å}$ als gute Abschätzung für die Trennung zwischen nahe- und nicht-nativem Bereich. Ohne Referenzstruktur stellte die Bewertung der Kandidatenkomplexe kritische Anforderungen an die eingesetzten Modelle. Sie mussten die Beteiligung aller Kontakte der molekularen Schnittstelle korrekt repräsentieren, ohne dabei in eine Abhängigkeit von der Größe ihrer Interaktionsfläche zu verfallen. Im Vergleich konnte diese Anforderungen nur die Bewertungsfunktion ITScore-PR erfüllen. Sie war damit in der Lage, native von nicht-nativen Kandidatenkomplexen zu unterscheiden. Da die Korrelation zwischen Bewertung und Nativität einer simulierten Komplexstruktur einem beträchtlichen Rauschen unterlag und auf den nahe-nativen Bereich beschränkt blieb, kann die korrekte Identifikation einzeln auftretender nahe-nativer Komplexe in einem kleinen Datensatz nicht unbedingt garantiert werden. Bei einer hinreichend großen Menge an Kandidatenstrukturen ist die Identifikation jedoch durch die höhere Abdeckung des nahe-nativen Bereichs deutlich verlässlicher. Ein fallübergreifender Vergleich der erzielten Bewertung ist nicht zweckmäßig. Da die Bewertungsfunktion ITScore-PR erfolgreich in der Lage war, die interatomaren Kontakte zwischen Protein und Nukleinsäure auch ohne Vorliegen einer Referenzstruktur zur Beurteilung der Nativität simulierter Komplexvarianten zu nutzen, kann die dritte Hypothese bestätigt werden.

Hypothese 4 Die bioinformatische Analyse erhobener Sequenzdaten erlaubt die Aufklärung der Bindekonstellation zwischen Aptamer und Zielprotein sowie die Erkennung bindungsrelevanter Sequenz- und Strukturmerkmale, die zum Entwurf einer optimierten Selektionsbibliothek eingesetzt werden können.

Nach Abschluss der methodischen Vorüberlegungen erfolgte die Anwendung der evaluierten Analysemethoden am Beispiel einer konkreten Aptamerselektion nach dem Verfahren SELEX. Der experimentelle Anteil wurde dabei von Mitarbeitern der TU Dresden geplant und durchgeführt. Um die Kapsidoberfläche des humanen Norovirus Genotyp `GII.4` mit einem Aptamer erkennen zu können, wurde dessen großes Kapsidprotein `VP1` in der Selektion als Zielmolekül eingesetzt. Mittels Hochdurchsatzsequenzierung wurde schließlich die Datengrundlage für die darauffolgenden Analysen geschaffen. Über den Verlauf der Aptamerselektion konnte sowohl durch die Bewertung ihrer Diversität als auch anhand der beobachteten Clusterbildung eine spezifische Anreicherung in der Selektionsbibliothek festgestellt werden. Für den häufigsten in der finalen Bibliothek befindlichen Aptamerkandidaten wurde zudem die Bindung zum Zielmolekül im Experiment bestätigt. Die Analyse der erhobenen Daten erfolgte mit den validierten Verfahren aus der Vorbetrachtung. Notwendige Kopplungen und Informationsflüsse zwischen den einzelnen Analyseschritten wurden mit dem Hauptaugenmerk auf Ergänzung und Validierung in einem Verfahrensprotokoll festgehalten. Für die Anwendung der auf n-Grammen basierten Regression mussten die relativen Affinitäten der Aptamerkandidaten zum Zielprotein anhand ihrer Häufigkeitsverteilung geschätzt werden. Das Verfahren ließ sich anschließend auf einen reduzierten Sequenzdatensatz aus der finalen Selektionsrunde anwenden und lieferte plausible Hinweise auf die Beteiligung spezifischer Sequenzbereiche der Länge 5 bis 6 an der Affinität der Aptamere. Durch die Festlegung strenger Begrenzungskriterien konnte der Mustersuchalgorithmus auf die vollen Datensätze von Aptamersequenzen mehrerer Runden,

darunter der finalen, angewendet werden. Die Ergebnisse der Mustersuche wiesen geringe Anteile von Variabilität auf, standen jedoch in prinzipieller Übereinstimmung mit den Ergebnissen der Regressionsanalyse. Es folgte die Bestimmung der Tertiärstrukturen, wobei für das experimentell validierte Aptamer eine mehrstufige molekulardynamische Simulation und für das Zielprotein eine komplexe Pipeline zur Homologiemodellierung eingesetzt wurde. Dabei zeigte sich, dass die im Rahmen der vorigen Analysen gefundene, bindungsrelevante Teilsequenz auf der Tertiärstruktur des Aptamers für das Zielprotein zugänglich gelegen war. Im Zuge einer Dockingsimulation wurden ohne Angabe einer Bindungsspezifikation Kandidatenkomplexe aus diesen beiden Partnern erzeugt. Ihre Bewertung erfolgte primär durch ITScore-PR, zu Vergleichs- und Validierungszwecken wurden jedoch weitere Bewertungsfunktionen hinzugezogen. Die Ergebnisse der kombinierten Analysen wiesen auf drei Bindemodi der beiden Komplexpartner hin, die sich aus der bereits identifizierten, bindungsrelevanten Teilsequenz des Aptamers und drei Epitopen des Zielproteins ergaben. Im assemblierten Viruskapsid ist nur einer dieser Bindemodi ohne sterische Behinderung möglich. Sowohl die Mustersuche, die Regression als auch die tertiärstrukturbasierte Bewertungsmethode stützten die Vermutung, dass das Sequenzmotiv `TGGTCCGG` des Aptamers wichtig für seine Bindung zum Zielprotein ist und daher für die Generierung einer optimierten Bibliothek eingesetzt werden kann. Da die bioinformatischen Analyseverfahren auf Basis der erhobenen Sequenzdaten in der Lage waren, die Bindekonstellation zwischen Aptamer und Zielprotein aufzuklären und dabei ein bindungsrelevantes Motiv mit Potential zur Optimierung der Selektionsbibliothek herauszustellen, kann die vierte Hypothese bestätigt werden.

7.2 Ausblick

Im Rahmen dieser Arbeit konnte der Wert der bioinformatischen Analyse in der post-experimentellen Phase der Aptamerselektion deutlich hervorgehoben werden. Durch die Kombination der Analysemethoden konnten nicht nur die drei Bindekonstellationen des Komplexes aus selektiertem Aptamer und dem als Zielmolekül eingesetzten einzelnen Kapsidprotein aufgeklärt werden. Von dieser Basis aus war auch der Schluss auf die praktische Einsatzfähigkeit des Aptamers in der Umgebung unversehrter Noroviruskapside möglich, obwohl bei diesen ein großer Teil der Bindeoberfläche des Kapsidproteins unzugänglich bleibt. Die Analyseergebnisse halfen ferner bei der Identifikation eines Sequenzmusters mit struktureller Einordnung, das für die Bindung zwischen Aptamer und Zielprotein wichtig ist und zur Verbesserung des Verfahrens eingesetzt werden kann. Da die grundlegende Thematik der Optimierung von Selektionsbibliotheken an dieser Stelle noch nicht abschließend behandelt ist, finden sich im Folgenden eine Reihe von Anknüpfungspunkten für eine inhaltliche Fortsetzung dieser Arbeit.

Evaluation von Analysemethoden Ein Anknüpfungspunkt ist die Weiterführung der Recherche und Evaluation von Analysemethoden, für die im Rahmen der vorliegenden Arbeit erste fundierte Grundlagen geschaffen wurden. So könnten neue Konzepte der Evaluation bisher unbeachtet gebliebene Aspekte der Methoden beleuchten und in die Wertung einbringen. Ein hoher Stellenwert sollte der Akquise qualitativ hochwertiger und hinreichend großer Datensätze zur Bildung eines Gold-Standards zugemessen werden. Dies umfasst hinsichtlich ihrer Affinität annotierte Aptamersequenzen genauso wie auch im größeren Umfang sequenzierte Bibliotheken der zugehörigen Selektionsexperimente, sowohl für ähnliche als auch unterschiedliche Zielproteine. Im Bereich der Tertiärstrukturen erlaubt die Verfügbarkeit der ungebundenen

Einzelstrukturen von Aptamer und Zielprotein sowie der Komplexstruktur beider darüber hinaus ein breiteres Spektrum der Betrachtung. Die stete methodische Weiterentwicklung führt ferner zur Notwendigkeit, neu aufkommende Methoden zu evaluieren. Diese könnten nicht nur die bisherig eingesetzten verbessern oder ergänzen, sondern auch neue Analysefelder eröffnen. Schließlich sei erwähnt, dass nur durch die Überprüfung der Verknüpfungen zwischen den Methoden und die anschließende Optimierung des Informationsflusses bei Änderungen sichergestellt werden kann, dass durch die methodische Kombination der maximale Gewinn erzielt wird. Insgesamt muss das Ziel sein, sowohl den Umfang als auch die Integrität der Evaluation weiter zu erhöhen.

Generierung optimierter Selektionsbibliotheken Die Ergebnisse der bioinformatischen Analysen stellen eine gute Basis für die Generierung von targetspezifisch optimierten Selektionsbibliotheken dar und bieten damit einen weiteren Anknüpfungspunkt an die vorliegende Arbeit. Aus den Analyseergebnissen lassen sich, wie bereits im Rahmen der Arbeit kurz umrissen wurde, unterschiedliche Unterräume des Sequenz- und Strukturraumes ableiten. Diese können sowohl direkt als auch in Kombination für die Erzeugung von Bibliotheken für darauffolgende Aptamerselektionen eingesetzt werden, müssen jedoch hinsichtlich ihres tatsächlichen Verbesserungspotentials untersucht werden. Nach gründlicher theoretischer Betrachtung sollten daher auf Basis der Erkenntnisse dieser Arbeit unterschiedliche Bibliotheken generiert, synthetisiert und im Selektionsexperiment mit dem bekannten Zielprotein `VP1` des humanen Norovirus evaluiert werden. Die Güte der Bibliotheken manifestiert sich nicht nur im Verlauf der Anreicherung innerhalb der Bibliothek sondern vor allem in der Sequenz und Affinität der häufigsten Aptamerkandidaten sowie im Vorhandensein bekannter und neuer Bindemotive. Von großem Interesse kann auch der Vergleich der Bindekonstellationen werden, die unter Einsatz verschiedener Bibliotheken erhalten wurden. Insgesamt ist mit diesem Vorgehen ein sehr hoher experimenteller und analytischer Aufwand verbunden. Bei sorgfältiger Bearbeitung können damit jedoch wichtige Erkenntnisse über die Anwendbarkeit der Optimierungskonzepte sowie über die Bedeutung und Wahl der Methodenparameter gewonnen werden, die den notwendigen Aufwand in nachfolgenden Fällen deutlich herabsetzen. Durch den Einsatz der optimierten Bibliotheken können schließlich leistungsfähigere Aptamere, also solche mit höherer Affinität oder Spezifität, für das Kapsid des humanen Norovirus gefunden werden.

Generalisierung des Verfahrens Mit der Durchführung und Analyse einer Aptamerselektion gegen das Kapsid des humanen Norovirus war der Fokus dieser Arbeit thematisch sehr spezifisch ausgerichtet. Das entwickelte Verfahrensprotokoll für die bioinformatische Analyse geht dabei zwar einen ersten Schritt in Richtung eines generalisierbaren Analyseverfahrens, seine praktische Bewährung an den Herausforderungen weiterer Anwendungsfälle steht jedoch noch aus. Weitere Studien können daher durch die Übertragung des in dieser Arbeit vorgeschlagenen Analyseverfahrens auf andere Zielproteine sinnvoll anknüpfen. Sie können dabei neben der initialen Aptamerselektion ohne Verwendung targetspezifischer *a priori*-Informationen und den anschließenden bioinformatischen Analysen auch die Erzeugung einer targetspezifisch optimierten Bibliothek einschließen, mit deren Hilfe schließlich leistungsfähigere Aptamere gewonnen werden können. Besonderer Wert sollte dabei auf die Identifikation der methodischen Grenzen gelegt werden, auf deren Basis alternative Methoden eingeführt und unnötige

Arbeitsschritte von vornherein vermieden werden können. Als langfristiges Ziel steht hierbei die Entwicklung eines zunehmend generalisierungsfähigen Analyseprotokolls, dessen Einsatz unabhängig vom Zielprotein der Aptamerselektion möglich ist.

Automatisierung des Verfahrens Da die bisher manuelle Durchführung und Koordination der Analysemethoden durch den verbundenen großen Aufwand eine nicht zu vernachlässigende Hürde für seine Akzeptanz in der wissenschaftlichen Gemeinschaft darstellt, kann im Fortgang die schrittweise Automatisierung des Verfahrensprotokolls forciert werden. Im ersten Schritt sollte dazu die programmatische Steuerung der Softwarekomponenten und des Informationsflusses durch die Einführung einheitlicher Schnittstellen für die einzelnen Analysemethoden gewährleistet werden. Großer Handlungsbedarf besteht anschließend bei der maschinellen Prozessierung der zahlreichen bis dahin menschlich getroffenen Entscheidungen. Diese müssen zunächst logisch nachvollziehbar verstanden sein und systematisiert werden, was Randbedingungen und mögliche Sonderfälle einschließt. Im Anschluss kann die Konzeption, Entwicklung und Testung entsprechender statistischer Entscheidungsmodelle auf dieser Basis erfolgen. Auch wenn eine vollständige Automatisierung an dieser Stelle nicht möglich ist, kann ein zunehmend selbstständiges, bioinformatisches Analysesystem mit minimaler Erfordernis für menschliches Eingreifen entstehen, welches im Fall der Intervention eine intuitive Benutzerführung bietet.

Literatur

[1] Rico Beier und Dirk Labudde. „Numeric promoter description – A comparative view on concepts and general application“. In: *Journal of Molecular Graphics and Modelling* 63 (Jan. 2016), S. 65–77. DOI: 10.1016/j.jmgm.2015.11.011 (siehe S. 24).

[2] Rico Beier, Elke Boschke und Dirk Labudde. „New Strategies for Evaluation and Analysis of SELEX Experiments“. In: *BioMed Research International* 2014 (2014), S. 1–12. DOI: 10.1155/2014/849743 (siehe S. 24).

[3] Rico Beier, Claudia Pahlke, Philipp Quenzel *et al.* „Selection of a DNA aptamer against norovirus capsid protein VP1“. In: *FEMS Microbiology Letters* 351.2 (Jan. 2014), S. 162–169. DOI: 10.1111/1574-6968.12366 (siehe S. 25).

[4] Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha und Alexey G. Murzin. „SCOP2 prototype: a new approach to protein structure mining“. In: *Nucleic Acids Research* 42.D1 (Nov. 2013), S. D310–D314. DOI: 10.1093/nar/gkt1242 (siehe S. 27).

[5] L Lins und R Brasseur. „The hydrophobic effect in protein folding.“ In: *The FASEB Journal* 9.7 (1995), S. 535–40 (siehe S. 27).

[6] C N Pace, B A Shirley, M McNutt und K Gajiwala. „Forces contributing to the conformational stability of proteins.“ In: *The FASEB Journal* 10.1 (1996), S. 75–83 (siehe S. 27).

[7] I. Schomburg, A. Chang, S. Placzek *et al.* „BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA“. In: *Nucleic Acids Research* 41.D1 (Nov. 2012), S. D764–D772. DOI: 10.1093/nar/gks1049 (siehe S. 27).

[8] C. F. Basler, A. Mikulasova, L. Martinez-Sobrido *et al.* „The Ebola Virus VP35 Protein Inhibits Activation of Interferon Regulatory Factor 3“. In: *Journal of Virology* 77.14 (Juli 2003), S. 7945–7956. DOI: 10.1128/jvi.77.14.7945-7956.2003 (siehe S. 27).

[9] Qing R. Fan und Wayne A. Hendrickson. „Structure of human follicle-stimulating hormone in complex with its receptor“. In: *Nature* 433.7023 (Jan. 2005), S. 269–277. DOI: 10.1038/nature03206 (siehe S. 27).

[10] Harry W. Schroeder und Lisa Cavacini. „Structure and function of immunoglobulins“. In: *Journal of Allergy and Clinical Immunology* 125.2 (Feb. 2010), S41–S52. DOI: 10.1016/j.jaci.2009.09.046 (siehe S. 27).

[11] P. K. Ferrigno. „Non-antibody protein-based biosensors“. In: *Essays In Biochemistry* 60.1 (Juni 2016), S. 19–25. DOI: 10.1042/ebc20150003 (siehe S. 27).

[12] Jennifer Audi, Martin Belson, Manish Patel, Joshua Schier und John Osterloh. „Ricin Poisoning“. In: *JAMA* 294.18 (Nov. 2005), S. 2342. DOI: 10.1001/jama.294.18.2342 (siehe S. 27).

[13] Sandra Tan, Hwee Tong Tan und Maxey C. M. Chung. „Membrane proteins and membrane proteomics“. In: *Proteomics* 8.19 (Okt. 2008), S. 3924–3932. DOI: 10.1002/pmic.200800597 (siehe S. 27).

[14] Roslyn M Bill, Peter J F Henderson, So Iwata *et al.* „Overcoming barriers to membrane protein structure determination“. In: *Nat Biotechnol* 29.4 (Apr. 2011), S. 335–340. DOI: 10.1038/nbt.1833 (siehe S. 27).

[15] J. Preben Morth, Bjørn P. Pedersen, Morten J. Buch-Pedersen *et al.* „A structural overview of the plasma membrane Na+, K+-ATPase and H+-ATPase ion pumps“. In: *Nature Reviews Molecular Cell Biology* 12.1 (Jan. 2011), S. 60–70. DOI: 10.1038/nrm3031 (siehe S. 28).

[16] Francesco Tombola, Medha M. Pathak und Ehud Y. Isacoff. „How Does Voltage Open an Ion Channel?“ In: *Annual Review of Cell and Developmental Biology* 22.1 (Nov. 2006), S. 23–52. DOI: 10.1146/annurev.cellbio.21.020404.145837 (siehe S. 28).

[17] David C. Gadsby. „Ion channels versus ion pumps: the principal difference, in principle“. In: *Nature Reviews Molecular Cell Biology* 10.5 (Apr. 2009), S. 344–352. DOI: 10.1038/nrm2668 (siehe S. 28).

[18] Charles Hachez, Arnaud Besserer, Adrien S. Chevalier und François Chaumont. „Insights into plant plasma membrane aquaporin trafficking“. In: *Trends in Plant Science* 18.6 (Juni 2013), S. 344–352. DOI: 10.1016/j.tplants.2012.12.003 (siehe S. 28).

[19] Daniel M. Rosenbaum, Søren G. F. Rasmussen und Brian K. Kobilka. „The structure and function of G-protein-coupled receptors“. In: *Nature* 459.7245 (Mai 2009), S. 356–363. DOI: 10.1038/nature08144 (siehe S. 28).

[20] Maurice L. Huggins. „The Structure of Fibrous Proteins.“ In: *Chemical Reviews* 32.2 (Apr. 1943), S. 195–218. DOI: 10.1021/cr60102a002 (siehe S. 28).

[21] Bin Wang, Wen Yang, Joanna McKittrick und Marc André Meyers. „Keratin: Structure, mechanical properties, occurrence in biological organisms, and efforts at bioinspiration“. In: *Progress in Materials Science* 76 (März 2016), S. 229–318. DOI: 10.1016/j.pmatsci.2015.06.001 (siehe S. 28).

[22] Vincent R. Sherman, Wen Yang und Marc A. Meyers. „The materials science of collagen“. In: *Journal of the Mechanical Behavior of Biomedical Materials* 52 (Dez. 2015), S. 22–50. DOI: 10.1016/j.jmbbm.2015.05.023 (siehe S. 28).

[23] Suzanne M. Mithieux und Anthony S. Weiss. „Elastin“. In: *Fibrous Proteins: Coiled-Coils, Collagen and Elastomers*. Elsevier BV, 2005, S. 437–461. DOI: 10.1016/s0065-3233(05)70013-9 (siehe S. 28).

[24] A PATEL, B FINE, M SANDIG und K MEQUANINT. „Elastin biosynthesis: The missing link in tissue-engineered blood vessels“. In: *Cardiovascular Research* 71.1 (Juli 2006), S. 40–49. DOI: 10.1016/j.cardiores.2006.02.021 (siehe S. 28).

[25] G. Steger. *Bioinformatik: Methoden Zur Vorhersage Von Rna- Und Proteinstrukturen*. Birkhäuser, 2003. ISBN: 9783764369514 (siehe S. 28–32, 35 f.).

[26] H. Hart, L.E. Craine und D.J. Hart. *Organische Chemie*. John Wiley & Sons, 2002. ISBN: 9783527303793 (siehe S. 28 f., 31, 34).

[27] D.P. Clark und A. Held. *Molecular Biology: Das Original Mit Übersetzungshilfen: Understanding the Genetic Revolution*. Easy-Reading. Spektrum, Akad. Verlag, 2006. ISBN: 9783827416964 (siehe S. 29–32).

[28] J. Koolman und K.H. Röhm. *TaschenAtlas der Biochemie*. Thieme flexibook. Thieme, 2003. ISBN: 9783137594031 (siehe S. 30).

[29] Russel J. Mortishire-Smith, Alex F. Drake, Jennifer C. Nutkins und Dudley H. Williams. „Left handed α-helix formation by a bacterial peptide“. In: *FEBS Letters* 278.2 (1991), S. 244–246. ISSN: 0014-5793. DOI: 10.1016/0014-5793(91)80126-N (siehe S. 30).

[30] Richard B. Cooley, Daniel J. Arp und P. Andrew Karplus. „Evolutionary Origin of a Secondary Structure: π-Helices as Cryptic but Widespread Insertional Variations of α-Helices That Enhance Protein Functionality“. In: *Journal of Molecular Biology* 404.2 (Nov. 2010), S. 232–246. DOI: 10.1016/j.jmb.2010.09.034 (siehe S. 30).

[31] MSOE Center for BioMolecular Modeling. *Protein Structure Tutorials. Secondary Structure: Alpha Helices and Beta Pleated Sheets*. URL: http://cbm.msoe.edu/teachingResources/proteinStructure/secondary.html (besucht am 18. 12. 2016) (siehe S. 31).

[32] James D. Watson, Tania A. Baker, Stephen P. Bell *et al. Molecular Biology of the Gene* -. 7th edition. München: Pearson, 2014. ISBN: 978-0-321-76243-6 (siehe S. 33–36).

[33] Jiří Šponer, Judit E. Šponer, Arnošt Mládek *et al.* „Nature and magnitude of aromatic base stacking in DNA and RNA: Quantum chemistry, molecular mechanics and experiment“. In: *Biopolymers* (Juni 2013), n/a–n/a. DOI: 10.1002/bip.22322 (siehe S. 35).

[34] K. Darty, A. Denise und Y. Ponty. „VARNA: Interactive drawing and editing of the RNA secondary structure“. In: *Bioinformatics* 25.15 (Apr. 2009), S. 1974–1975. DOI: 10.1093/bioinformatics/btp250 (siehe S. 35 f., 178).

[35] Robert T. Batey, Robert P. Rambo und Jennifer A. Doudna. „Tertiary Motifs in RNA Structure and Folding“. In: *Angewandte Chemie International Edition* 38.16 (1999), S. 2326–2343. ISSN: 1521-3773. DOI: 10.1002/(SICI)1521-3773(19990816)38:16<2326::AID-ANIE2326>3.0.CO;2-3 (siehe S. 36).

[36] Phoebe A. Rice und Carl C. Correll. *Protein-Nucleic Acid Interactions - Structural Biology*. 2008. Aufl. Cambridge: Royal Society of Chemistry, 2008. ISBN: 978-0-854-04272-2 (siehe S. 36 f.).

[37] N. M. Luscombe. „Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level“. In: *Nucleic Acids Research* 29.13 (Juli 2001), S. 2860–2874. DOI: 10.1093/nar/29.13.2860 (siehe S. 36 f.).

[38] K. A. Wilson, J. L. Kellie und S. D. Wetmore. „DNA-protein -interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar“. In: *Nucleic Acids Research* 42.10 (Apr. 2014), S. 6726–6741. DOI: 10.1093/nar/gku269 (siehe S. 36).

[39] Lesley R. Rutledge, Lachlan S. Campbell-Verduyn und Stacey D. Wetmore. „Characterization of the stacking interactions between DNA or RNA nucleobases and the aromatic amino acids“. In: *Chemical Physics Letters* 444.1-3 (Aug. 2007), S. 167–175. DOI: 10.1016/j.cplett.2007.06.090 (siehe S. 36).

[40] Remo Rohs, Xiangshu Jin, Sean M. West *et al.* „Origins of Specificity in Protein-DNA Recognition“. In: *Annu. Rev. Biochem.* 79.1 (Juni 2010), S. 233–269. DOI: 10.1146/annurev-biochem-060408-091030 (siehe S. 36).

[41] B. TURNER. „Induced fit of RNA on binding the L7Ae protein to the kink-turn motif“. In: *RNA* 11.8 (Aug. 2005), S. 1192–1200. DOI: 10.1261/rna.2680605 (siehe S. 37).

[42] Lizhe Zhu, Hanlun Jiang, Fu Kit Sheong *et al.* „A Flexible Domain-Domain Hinge Promotes an Induced-fit Dominant Mechanism for the Loading of Guide-DNA into Argonaute Protein in Thermus thermophilus“. In: *The Journal of Physical Chemistry B* 120.10 (März 2016), S. 2709–2720. DOI: 10.1021/acs.jpcb.5b12426 (siehe S. 37).

[43] Aditi Gupta und Michael Gribskov. „The Role of RNA Sequence and Structure in RNA-Protein Interactions“. In: *Journal of Molecular Biology* 409.4 (Juni 2011), S. 574–587. DOI: 10.1016/j.jmb.2011.04.007 (siehe S. 37).

[44] Laura Pérez-Cano und Juan Fernández-Recio. „Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites on proteins“. In: *Proteins* 78.1 (Juli 2009), S. 25–35. DOI: 10.1002/prot.22527 (siehe S. 38, 102).

[45] David E Draper. „Themes in RNA-protein recognition“. In: *Journal of Molecular Biology* 293.2 (1999), S. 255–270. ISSN: 0022-2836. DOI: 10.1006/jmbi.1999.2991 (siehe S. 38).

[46] Agnes Noy, Alberto Pérez, Filip Lankas, F. Javier Luque und Modesto Orozco. „Relative Flexibility of DNA and RNA: a Molecular Dynamics Study“. In: *Journal of Molecular Biology* 343.3 (Okt. 2004), S. 627–638. DOI: 10.1016/j.jmb.2004.07.048 (siehe S. 38).

[47] A. Perez. „The relative flexibility of B-DNA and A-RNA duplexes: database analysis“. In: *Nucleic Acids Research* 32.20 (Nov. 2004), S. 6144–6151. DOI: 10.1093/nar/gkh954 (siehe S. 38).

[48] Regina Stoltenburg, Christine Reinemann und Beate Strehlitz. „SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands“. In: *Biomolecular Engineering* 24.4 (Okt. 2007), S. 381–403. DOI: 10.1016/j.bioeng.2007.06.001 (siehe S. 39 f., 42, 44–48, 144).

[49] T. Hermann. „Adaptive Recognition by Nucleic Acid Aptamers“. In: *Science* 287.5454 (Feb. 2000), S. 820–825. DOI: 10.1126/science.287.5454.820 (siehe S. 39).

[50] Hao Qu, Andrew T. Csordas, Jinpeng Wang *et al.* „Rapid and Label-Free Strategy to Isolate Aptamers for Metal Ions“. In: *ACS Nano* 10.8 (Aug. 2016), S. 7558–7565. DOI: 10.1021/acsnano.6b02558 (siehe S. 40).

[51] Jan Wrzesinski und Jerzy Ciesiolka. „Characterization of Structure and Metal Ions Specificity of Co 2+ -Binding RNA Aptamers †“. In: *Biochemistry* 44.16 (Apr. 2005), S. 6257–6268. DOI: 10.1021/bi047397u (siehe S. 40).

[52] H. P. Hofmann, S. Limmer, V. Hornung und M. Sprinzl. „Ni2+-binding RNA motifs with an asymmetric purine-rich internal loop and a G-A base pair“. In: *RNA* 3.11 (Nov. 1997), S. 1289–1300 (siehe S. 40).

[53] J. Ciesiolka, J. Gorski und M. Yarus. „Selection of an RNA domain that binds Zn2+“. In: *RNA* 1.5 (Juli 1995), S. 538–550 (siehe S. 40).

[54] Jijun Tang, Jianwei Xie, Ningsheng Shao und Yan Yan. „The DNA aptamers that specifically recognize ricin toxin are selected by twoin vitro selection methods“. In: *ELECTROPHORESIS* 27.7 (Apr. 2006), S. 1303–1311. DOI: 10.1002/elps.200500489 (siehe S. 40, 46).

[55] Shimaa Eissa, Mohamed Siaj und Mohammed Zourob. „Aptamer-based competitive electrochemical biosensor for brevetoxin-2“. In: *Biosensors and Bioelectronics* 69 (Juli 2015), S. 148–154. DOI: 10.1016/j.bios.2015.01.055 (siehe S. 40).

[56] Libing Wang, Wenwei Ma, Wei Chen *et al.* „An aptamer-based chromatographic strip assay for sensitive toxin semi-quantitative detection“. In: *Biosensors and Bioelectronics* 26.6 (Feb. 2011), S. 3059–3062. DOI: 10.1016/j.bios.2010.11.040 (siehe S. 40).

[57] Cecilia Mannironi, Alessia Di Nardo, Paolo Fruscoloni und G. P. Tocchini-Valentini. „In Vitro Selection of Dopamine RNA Ligands“. In: *Biochemistry* 36.32 (Aug. 1997), S. 9726–9734. DOI: 10.1021/bi9700633 (siehe S. 40).

[58] Dominique Lévesque, Jean-Denis Beaudoin, Sébastien Roy und Jean-Pierre Perreault. „In vitro selection and characterization of RNA aptamers binding thyroxine hormone“. In: *Biochem. J.* 403.1 (Apr. 2007), S. 129–138. DOI: 10.1042/bj20061216 (siehe S. 40).

[59] Minjoung Jo, Ji-Young Ahn, Joohyung Lee *et al.* „Development of Single-Stranded DNA Aptamers for Specific Bisphenol A Detection“. In: *Oligonucleotides* 21.2 (Apr. 2011), S. 85–91. DOI: 10.1089/oli.2010.0267 (siehe S. 40).

[60] Daiying Xu, Vamsee-Krishna Chatakonda, Antonis Kourtidis, Douglas S. Conklin und Hua Shi. „In Search of Novel Drug Target Sites on Estrogen Receptors Using RNA Aptamers“. In: *Nucleic Acid Therapeutics* 24.3 (Juni 2014), S. 226–238. DOI: 10.1089/nat.2013.0474 (siehe S. 40).

[61] Milan N. Stojanovic, Paloma de Prada und Donald W. Landry. „Fluorescent Sensors Based on Aptamer Self-Assembly“. In: *J. Am. Chem. Soc.* 122.46 (Nov. 2000), S. 11547–11548. DOI: 10.1021/ja0022223 (siehe S. 40).

[62] Mohsen Ebrahimi, Hossein Hamzeiy, Jaleh Barar, Abolfazl Barzegari und Yadollah Omidi. „Systematic Evolution of Ligands by Exponential Enrichment Selection of Specific Aptamer for Sensing of Methamphetamine“. In: *Sensor Letters* 11.3 (März 2013), S. 566–570. DOI: 10.1166/sl.2013.2824 (siehe S. 40).

[63] Dilara Grate und Charles Wilson. „Inducible regulation of the S. cerevisiae cell cycle mediated by an RNA aptamer–ligand complex“. In: *Bioorganic & Medicinal Chemistry* 9.10 (Okt. 2001), S. 2565–2570. DOI: 10.1016/s0968-0896(01)00031-1 (siehe S. 40).

[64] Jeremy R. Babendure, Stephen R. Adams und Roger Y. Tsien. „Aptamers Switch on Fluorescence of Triphenylmethane Dyes“. In: *J. Am. Chem. Soc.* 125.48 (Dez. 2003), S. 14716–14717. DOI: 10.1021/ja037994o (siehe S. 40).

[65] Charles Wilson und Jack W. Szostak. „Isolation of a fluorophore-specific DNA aptamer with weak redox activity“. In: *Chemistry & Biology* 5.11 (Nov. 1998), S. 609–617. DOI: 10.1016/s1074-5521(98)90289-7 (siehe S. 40).

[66] Doerthe Mann, Christine Reinemann, Regina Stoltenburg und Beate Strehlitz. „In vitro selection of DNA aptamers binding ethanolamine“. In: *Biochemical and Biophysical Research Communications* 338.4 (Dez. 2005), S. 1928–1934. DOI: 10.1016/j.bbrc.2005.10.172 (siehe S. 40).

[67] R. Jenison, S. Gill, A Pardi und B Polisky. „High-resolution molecular discrimination by RNA“. In: *Science* 263.5152 (März 1994), S. 1425–1429. DOI: 10.1126/science.7510417 (siehe S. 40, 47).

[68] Atsushi Okazawa, Hiroshi Maeda, Eiichiro Fukusaki, Yoshio Katakura und Akio Kobayashi. „In vitro selection of hematoporphyrin binding DNA aptamers“. In: *Bioorganic & Medicinal Chemistry Letters* 10.23 (Dez. 2000), S. 2653–2656. DOI: 10.1016/s0960-894x(00)00540-0 (siehe S. 40).

[69] Teru Kato, Taro Takemura, Kazuyoshi Yano, Kazunori Ikebukuro und Isao Karube. „In vitro selection of DNA aptamers which bind to cholic acid“. In: *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1493.1-2 (Sep. 2000), S. 12–18. DOI: 10.1016/s0167-4781(00)00080-4 (siehe S. 40).

[70] D. Kiga, Y. Futamura, K. Sakamoto und S. Yokoyama. „An RNA aptamer to the xanthine/-guanine base with a distinctive mode of purine recognition“. In: *Nucleic Acids Research* 26.7 (Apr. 1998), S. 1755–1760. DOI: 10.1093/nar/26.7.1755 (siehe S. 40).

[71] Marc Meli, Jacques Vergne, Jean-Luc Décout und Marie-Christine Maurel. „Adenine-Aptamer Complexes“. In: *Journal of Biological Chemistry* 277.3 (Nov. 2001), S. 2104–2111. DOI: 10.1074/jbc.m107130200 (siehe S. 40).

[72] Dalila Sekkai, Eric Dausse, Carmelo Di Primo *et al.* „In Vitro Selection of DNA Aptamers Against the HIV-1 TAR RNA Hairpin“. In: *Antisense and Nucleic Acid Drug Development* 12.4 (Aug. 2002), S. 265–274. DOI: 10.1089/108729002320351584 (siehe S. 40).

[73] D. Scarabino. „tRNA prefers to kiss“. In: *The EMBO Journal* 18.16 (Aug. 1999), S. 4571–4578. DOI: 10.1093/emboj/18.16.4571 (siehe S. 40).

[74] Sunjoo Jeong, Tae-Yeon Eom, Se-Jin Kim, Seong-Wook Lee und Jaehoon Yu. „In Vitro Selection of the RNA Aptamer against the Sialyl Lewis X and Its Inhibition of the Cell Adhesion“. In: *Biochemical and Biophysical Research Communications* 281.1 (Feb. 2001), S. 237–243. DOI: 10.1006/bbrc.2001.4327 (siehe S. 40).

[75] Ei-ichiro Fukusaki, Takahisa Kato, Hiroshi Maeda *et al.* „DNA aptamers that bind to chitin“. In: *Bioorganic & Medicinal Chemistry Letters* 10.5 (März 2000), S. 423–425. DOI: 10.1016/s0960-894x(00)00013-5 (siehe S. 40).

[76] B. J. Boese und R. R. Breaker. „In vitro selection and characterization of cellulose-binding DNA aptamers“. In: *Nucleic Acids Research* 35.19 (Okt. 2007), S. 6378–6388. DOI: 10.1093/nar/gkm708 (siehe S. 40).

[77] M. Legiewicz. „A More Complex Isoleucine Aptamer with a Cognate Triplet“. In: *Journal of Biological Chemistry* 280.20 (März 2005), S. 19815–19822. DOI: 10.1074/jbc.m502329200 (siehe S. 40).

[78] I. Majerfeld. „A diminutive and specific RNA binding site for L-tryptophan“. In: *Nucleic Acids Research* 33.17 (Sep. 2005), S. 5482–5493. DOI: 10.1093/nar/gki861 (siehe S. 40).

[79] J RUTA, C GROSSET, C RAVELET *et al.* „Chiral resolution of histidine using an anti-d-histidine l-RNA aptamer microbore column“. In: *Journal of Chromatography B* 845.2 (Jan. 2007), S. 186–190. DOI: 10.1016/j.jchromb.2006.06.026 (siehe S. 40).

[80] Yoshihiko Nonaka, Wataru Yoshida, Koichi Abe *et al.* „Affinity Improvement of a VEGF Aptamer by in Silico Maturation for a Sensitive VEGF-Detection System“. In: *Analytical Chemistry* 85.2 (Jan. 2013), S. 1132–1137. DOI: 10.1021/ac303023d (siehe S. 40).

[81] Louis C. Bock, Linda C. Griffin, John A. Latham, Eric H. Vermaas und John J. Toole. „Selection of single-stranded DNA molecules that bind and inhibit human thrombin“. In: *Nature* 355.6360 (Feb. 1992), S. 564–566. DOI: 10.1038/355564a0 (siehe S. 40).

[82] Shaun D. Mendonsa und Michael T. Bowser. „In Vitro Selection of High-Affinity DNA Ligands for Human IgE Using Capillary Electrophoresis“. In: *Analytical Chemistry* 76.18 (Sep. 2004), S. 5387–5392. DOI: 10.1021/ac049857v (siehe S. 40).

[83] L. A. Jones, L. E. Clancy, W. D. Rawlinson und P. A. White. „High-Affinity Aptamers to Subtype 3a Hepatitis C Virus Polymerase Display Genotypic Specificity“. In: *Antimicrobial Agents and Chemotherapy* 50.9 (Aug. 2006), S. 3019–3027. DOI: 10.1128/aac.01603-05 (siehe S. 40).

[84] C.S.M. Ferreira, C.S. Matthews und S. Missailidis. „DNA Aptamers That Bind to MUC1 Tumour Marker: Design and Characterization of MUC1-Binding Single-Stranded DNA Aptamers“. In: *Tumor Biology* 27.6 (2006), S. 289–301. DOI: 10.1159/000096085 (siehe S. 40).

[85] Petra Burgstaller und Michael Famulok. „Isolation of RNA Aptamers for Biological Cofactors by In Vitro Selection“. In: *Angewandte Chemie International Edition in English* 33.10 (Juni 1994), S. 1084–1087. DOI: 10.1002/anie.199410841 (siehe S. 40).

[86] Charles T. Lauhon und Jack W. Szostak. „RNA aptamers that bind flavin and nicotinamide redox cofactors“. In: *J. Am. Chem. Soc.* 117.4 (Feb. 1995), S. 1246–1257. DOI: 10.1021/ja00109a008 (siehe S. 40).

[87] Nadia Nikolaus und Beate Strehlitz. „DNA-Aptamers Binding Aminoglycoside Antibiotics“. In: *Sensors* 14.2 (Feb. 2014), S. 3737–3755. DOI: 10.3390/s140203737 (siehe S. 40).

[88] Heike Schürer, Katherina Stembera, Dietmar Knoll *et al.* „Aptamers that bind to the antibiotic moenomycin A“. In: *Bioorganic & Medicinal Chemistry* 9.10 (Okt. 2001), S. 2557–2563. DOI: 10.1016/s0968-0896(01)00030-x (siehe S. 40).

[89] Javed H. Niazi, Su Jin Lee und Man Bock Gu. „Single-stranded DNA aptamers specific for antibiotics tetracyclines“. In: *Bioorganic & Medicinal Chemistry* 16.15 (Aug. 2008), S. 7245–7253. DOI: 10.1016/j.bmc.2008.06.033 (siehe S. 40).

[90] Jee-Woong Park, Su Jin Lee, Eun-Jin Choi *et al.* „An ultra-sensitive detection of a whole virus using dual aptamers developed by immobilization-free screening“. In: *Biosensors and Bioelectronics* 51 (Jan. 2014), S. 324–329. DOI: 10.1016/j.bios.2013.07.052 (siehe S. 40).

[91] Andreas Nitsche, Andreas Kurth, Anna Dunkhorst *et al.* „One-step selection of Vaccinia virus-binding DNA aptamers by MonoLEX“. In: *BMC Biotechnology* 7.1 (2007), S. 48. DOI: 10.1186/1472-6750-7-48 (siehe S. 40, 49).

[92] Milada Ikanovic, Walter E. Rudzinski, John G. Bruno *et al.* „Fluorescence Assay Based on Aptamer-Quantum Dot Binding to Bacillus thuringiensis Spores“. In: *J Fluoresc* 17.2 (Jan. 2007), S. 193–199. DOI: 10.1007/s10895-007-0158-4 (siehe S. 40).

[93] John G Bruno und Johnathan L Kiel. „In vitro selection of DNA aptamers to anthrax spores with electrochemiluminescence detection“. In: *Biosensors and Bioelectronics* 14.5 (Mai 1999), S. 457–464. DOI: 10.1016/s0956-5663(99)00028-7 (siehe S. 40).

[94] Zhiwen Tang, Dihua Shangguan, Kemin Wang *et al.* „Selection of Aptamers for Molecular Recognition and Characterization of Cancer Cells“. In: *Analytical Chemistry* 79.13 (Juli 2007), S. 4900–4907. DOI: 10.1021/ac070189y (siehe S. 40).

[95] Hari P. Dwivedi, R. Derike Smiley und Lee-Ann Jaykus. „Selection and characterization of DNA aptamers with binding selectivity to Campylobacter jejuni using whole-cell SELEX“. In: *Appl Microbiol Biotechnol* 87.6 (Juni 2010), S. 2323–2334. DOI: 10.1007/s00253-010-2728-7 (siehe S. 40).

[96] M. Homann und H. U. Goringer. „Combinatorial selection of high affinity RNA ligands to live African trypanosomes“. In: *Nucleic Acids Research* 27.9 (Mai 1999), S. 2006–2014. DOI: 10.1093/nar/27.9.2006 (siehe S. 40).

[97] Camille L. A. Hamula, Hongquan Zhang, Le Luo Guan, Xing-Fang Li und X. Chris Le. „Selection of Aptamers against Live Bacterial Cells“. In: *Analytical Chemistry* 80.20 (Okt. 2008), S. 7812–7819. DOI: 10.1021/ac801272s (siehe S. 40).

[98] Sonia Amaya-González, Noemí de-los-Santos-Álvarez, Arturo J. Miranda-Ordieres und M. Jesús Lobo-Castañón. „Aptamer Binding to Celiac Disease-Triggering Hydrophobic Proteins: A Sensitive Gluten Detection Approach“. In: *Analytical Chemistry* 86.5 (März 2014), S. 2733–2739. DOI: 10.1021/ac404151n (siehe S. 40, 48).

[99] N. Said, R. Rieder, R. Hurwitz *et al.* „In vivo expression and purification of aptamer-tagged small RNA regulators“. In: *Nucleic Acids Research* 37.20 (Sep. 2009), e133–e133. DOI: 10.1093/nar/gkp719 (siehe S. 40).

[100] Sanjay Tyagi. „Imaging intracellular RNA distribution and dynamics in living cells“. In: *Nature Methods* 6.5 (Mai 2009), S. 331–338. DOI: 10.1038/nmeth.1321 (siehe S. 40).

[101] Hideyuki Terazono, Yu Anzai, Mikhail Soloviev und Kenji Yasuda. „Labelling of live cells using fluorescent aptamers: binding reversal with DNA nucleases“. In: *Journal of Nanobiotechnology* 8.1 (2010), S. 8. DOI: 10.1186/1477-3155-8-8 (siehe S. 40).

[102] Wenjuan Wang, Chunlai Chen, Minxie Qian und Xin Sheng Zhao. „Aptamer biosensor for protein detection using gold nanoparticles“. In: *Analytical Biochemistry* 373.2 (Feb. 2008), S. 213–219. DOI: 10.1016/j.ab.2007.11.013 (siehe S. 40).

[103] Ming Li, Jianming Zhang, Savan Suri *et al.* „Detection of Adenosine Triphosphate with an Aptamer Biosensor Based on Surface-Enhanced Raman Scattering“. In: *Analytical Chemistry* 84.6 (März 2012), S. 2837–2842. DOI: 10.1021/ac203325z (siehe S. 40).

[104] Seram Lee, Young Sook Kim, Minjung Jo *et al.* „Chip-based detection of hepatitis C virus using RNA aptamers that specifically bind to HCV core antigen“. In: *Biochemical and Biophysical Research Communications* 358.1 (Juni 2007), S. 47–52. DOI: 10.1016/j.bbrc.2007.04.057 (siehe S. 40).

[105] A. Tahiri-Alaoui. „High affinity nucleic acid aptamers for streptavidin incorporated into bi-specific capture ligands“. In: *Nucleic Acids Research* 30.10 (Mai 2002), 45e–45. DOI: 10.1093/nar/30.10.e45 (siehe S. 40).

[106] Adam C. Connor und Linda B. McGown. „Aptamer stationary phase for protein capture in affinity capillary chromatography“. In: *Journal of Chromatography A* 1111.2 (Apr. 2006), S. 115–119. DOI: 10.1016/j.chroma.2005.05.012 (siehe S. 40).

[107] J.-H. Lee, M. D. Canny, A. De Erkenez *et al.* „A therapeutic aptamer inhibits angiogenesis by specifically targeting the heparin binding domain of VEGF165“. In: *Proceedings of the National Academy of Sciences* 102.52 (Dez. 2005), S. 18902–18907. DOI: 10.1073/pnas.0509069102 (siehe S. 40).

[108] Subash C. B. Gopinath, Yasuo Shikamoto, Hiroshi Mizuno und Penmetcha K. R. Kumar. „A potent anti-coagulant RNA aptamer inhibits blood coagulation by specifically blocking the extrinsic clotting pathway“. In: *Thromb Haemost* (Apr. 2006). DOI: 10.1160/th06-01-0047 (siehe S. 40).

[109] Yun-Hee Kim, Ho Jin Sung, Sukyoung Kim *et al.* „An RNA aptamer that specifically binds pancreatic adenocarcinoma up-regulated factor inhibits migration and growth of pancreatic cancer cells“. In: *Cancer Letters* 313.1 (Dez. 2011), S. 76–83. DOI: 10.1016/j.canlet.2011.08.027 (siehe S. 40).

[110] Claudia M. Dollins, Smita Nair, David Boczkowski *et al.* „Assembling OX40 Aptamers on a Molecular Scaffold to Create a Receptor-Activating Aptamer“. In: *Chemistry & Biology* 15.7 (Juli 2008), S. 675–682. DOI: 10.1016/j.chembiol.2008.05.016 (siehe S. 41).

[111] Elizabeth D. Pratico, Bruce A. Sullenger und Smita K. Nair. „Identification and Characterization of an Agonistic Aptamer Against the T Cell Costimulatory Receptor, OX40“. In: *Nucleic Acid Therapeutics* 23.1 (Feb. 2013), S. 35–43. DOI: 10.1089/nat.2012.0388 (siehe S. 41).

[112] Katarzyna Pala, Anna Serwotka, Filip Jeleń, Piotr Jakimowicz und Jacek Otlewski. „Tumor-specific hyperthermia with aptamer-tagged superparamagnetic nanoparticles“. In: *International Journal of Nanomedicine* (Dez. 2013), S. 67. DOI: 10.2147/ijn.s52539 (siehe S. 41).

[113] Yu-Fen Huang, Dihua Shangguan, Haipeng Liu *et al.* „Molecular Assembly of an Aptamer-Drug Conjugate for Targeted Drug Delivery to Tumor Cells“. In: *ChemBioChem* 10.5 (März 2009), S. 862–868. DOI: 10.1002/cbic.200800805 (siehe S. 41).

[114] Congsheng Cheng, Yong Hong Chen, Kim A Lennox, Mark A Behlke und Beverly L Davidson. „In vivo SELEX for Identification of Brain-penetrating Aptamers“. In: *Molecular Therapy — Nucleic Acids* 2.1 (Jan. 2013), e67. DOI: 10.1038/mtna.2012.59 (siehe S. 41).

[115] Jagat R. Kanwar, Kislay Roy, Nihal G. Maremanda *et al.* „Nucleic Acid-Based Aptamers: Applications, Development and Clinical Trials“. In: *Current Medicinal Chemistry* 22.21 (1. Juli 2015), S. 2539–2557. ISSN: 0929-8673 (siehe S. 41 ff.).

[116] Yeh-Hsing Lao, Kyle K.L. Phua und Kam W. Leong. „Aptamer Nanomedicine for Cancer Therapeutics: Barriers and Potential for Translation“. In: *ACS Nano* 9.3 (März 2015), S. 2235–2254. DOI: 10.1021/nn507494p (siehe S. 41).

[117] Suzy Kedzierski, Thomas Caltagirone und Makan Khoshnejad. „Synthetic Antibodies: The Emerging Field of Aptamers“. In: *BioProcessing Journal* 11.4 (Jan. 2013), S. 46–49. DOI: 10.12665/j114.kedzierskicaltagirone (siehe S. 42).

[118] Günter Mayer. „The Chemical Biology of Aptamers“. In: *Angewandte Chemie International Edition* 48.15 (März 2009), S. 2672–2689. DOI: 10.1002/anie.200804643 (siehe S. 43).

[119] Carlos Briones und Miguel Moreno. „Applications of peptide nucleic acids (PNAs) and locked nucleic acids (LNAs) in biosensor development“. In: *Anal Bioanal Chem* 402.10 (Feb. 2012), S. 3071–3089. DOI: 10.1007/s00216-012-5742-z (siehe S. 43).

[120] Axel Vater und Sven Klussmann. „Turning mirror-image oligonucleotides into drugs: the evolution of Spiegelmer® therapeutics“. In: *Drug Discovery Today* 20.1 (Jan. 2015), S. 147–155. DOI: 10.1016/j.drudis.2014.09.004 (siehe S. 43).

[121] C Tuerk und L Gold. „Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase“. In: *Science* 249.4968 (Aug. 1990), S. 505–510. DOI: 10.1126/science.2200121 (siehe S. 43).

[122] Mayumi Takahashi, Xiwei Wu, Michelle Ho *et al.* „High throughput sequencing analysis of RNA libraries reveals the influences of initial library and PCR methods on SELEX efficiency“. In: *Scientific Reports* 6 (Sep. 2016), S. 33697. DOI: 10.1038/srep33697 (siehe S. 45).

[123] Subash Chandra Bose Gopinath. „Methods developed for SELEX“. In: *Anal Bioanal Chem* 387.1 (Okt. 2006), S. 171–182. DOI: 10.1007/s00216-006-0826-2 (siehe S. 46).

[124] J. Liu. „Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions“. In: *Nucleic Acids Research* 33.17 (Sep. 2005), e141–e141. DOI: 10.1093/nar/gni139 (siehe S. 46).

[125] R. Stoltenburg, C. Reinemann und B. Strehlitz. „FluMag-SELEX as an advantageous method for DNA aptamer selection“. In: *Anal Bioanal Chem* 383.1 (Juli 2005), S. 83–91. DOI: 10.1007/s00216-005-3388-9 (siehe S. 46).

[126] M. Blank, T. Weinschenk, M. Priemer und H. Schluesener. „Systematic Evolution of a DNA Aptamer Binding to Rat Brain Tumor Microvessels: SELECTIVE TARGETING OF ENDOTHELIAL REGULATORY PROTEIN PIGPEN“. In: *Journal of Biological Chemistry* 276.19 (Feb. 2001), S. 16464–16468. DOI: 10.1074/jbc.m100347200 (siehe S. 46).

[127] Tomoko S. Misono und Penmetcha K.R. Kumar. „Selection of RNA aptamers against human influenza virus hemagglutinin using surface plasmon resonance“. In: *Analytical Biochemistry* 342.2 (Juli 2005), S. 312–317. DOI: 10.1016/j.ab.2005.04.013 (siehe S. 46).

[128] Susanne Meyer, John P. Maufort, Jeff Nie *et al.* „Development of an Efficient Targeted Cell-SELEX Procedure for DNA Aptamer Reagents“. In: *PLoS ONE* 8.8 (Aug. 2013). Hrsg. von Gangjian Qin, e71798. DOI: 10.1371/journal.pone.0071798 (siehe S. 46).

[129] S. C. B. Gopinath. „An RNA aptamer that distinguishes between closely related human influenza viruses and inhibits haemagglutinin-mediated membrane fusion“. In: *Journal of General Virology* 87.3 (März 2006), S. 479–487. DOI: 10.1099/vir.0.81508-0 (siehe S. 46).

[130] Rasa Beinoravičiūtė-Kellner, Georg Lipps und Gerhard Krauss. „In vitro selection of DNA binding sites for ABF1 protein from Saccharomyces cerevisiae“. In: *FEBS Letters* 579.20 (Juli 2005), S. 4535–4540. DOI: 10.1016/j.febslet.2005.07.009 (siehe S. 46).

[131] Tatjana Schütze, Barbara Wilhelm, Nicole Greiner *et al.* „Probing the SELEX Process with Next-Generation Sequencing“. In: *PLoS ONE* 6.12 (Dez. 2011). Hrsg. von Jörg D. Hoheisel, e29604. DOI: 10.1371/journal.pone.0029604 (siehe S. 46 f.).

[132] Michele Bianchini, Martín Radrizzani, Mariana G. Brocardo *et al.* „Specific oligobodies against ERK-2 that recognize both the native and the denatured state of the protein“. In: *Journal of Immunological Methods* 252.1-2 (Juni 2001), S. 191–197. DOI: 10.1016/s0022-1759(01)00350-7 (siehe S. 46, 48).

[133] S. Weiss, D. Proske, M. Neumann *et al.* „RNA aptamers specifically interact with the prion protein PrP“. In: *J. Virol.* 71.11 (Nov. 1997), S. 8790–8797 (siehe S. 46).

[134] M. G. Theis, A. Knorre, B. Kellersch *et al.* „Discriminatory aptamer reveals serum response element transcription regulated by cytohesin-2“. In: *Proceedings of the National Academy of Sciences* 101.31 (Juli 2004), S. 11221–11226. DOI: 10.1073/pnas.0402901101 (siehe S. 46).

[135] PHILIPPE BRIDONNEAU, YING-FON CHANG, ADA VELATI-BELLINI BUVOLI, DAN O'CONNELL und DAVID PARMA. „Site-Directed Selection of Oligonucleotide Antagonists by Competitive Elution". In: *Antisense and Nucleic Acid Drug Development* 9.1 (Feb. 1999), S. 1–11. DOI: 10.1089/oli.1.1999.9.1 (siehe S. 46).

[136] M. Svobodová, A. Pinto, P. Nadal und C. K. O' Sullivan. „Comparison of different methods for generation of single-stranded DNA for SELEX processes". In: *Anal Bioanal Chem* 404.3 (Juni 2012), S. 835–842. DOI: 10.1007/s00216-012-6183-4 (siehe S. 46).

[137] Marko Djordjevic. „SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways". In: *Biomolecular Engineering* 24.2 (Juni 2007), S. 179–189. DOI: 10.1016/j.bioeng.2007.03.001 (siehe S. 46, 48).

[138] Vincent J. B. Ruigrok, Mark Levisson, Johan Hekelaar *et al.* „Characterization of Aptamer-Protein Complexes by X-ray Crystallography and Alternative Approaches". In: *International Journal of Molecular Sciences* 13.12 (Aug. 2012), S. 10537–10552. DOI: 10.3390/ijms130810537 (siehe S. 47).

[139] Andrew D. Ellington und Jack W. Szostak. „Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures". In: *Nature* 355.6363 (Feb. 1992), S. 850–852. DOI: 10.1038/355850a0 (siehe S. 47).

[140] Chenglong Wang, Ming Zhang, Guang Yang *et al.* „Single-stranded DNA aptamers that bind differentiated but not parental cells: subtractive systematic evolution of ligands by exponential enrichment". In: *Journal of Biotechnology* 102.1 (Apr. 2003), S. 15–22. DOI: 10.1016/s0168-1656(02)00360-7 (siehe S. 48).

[141] Wei WANG und Ling-Yun JIA. „Progress in Aptamer Screening Methods". In: *Chinese Journal of Analytical Chemistry* 37.3 (März 2009), S. 454–460. DOI: 10.1016/s1872-2040(08)60092-4 (siehe S. 48 f.).

[142] Youngmi Kim, Chen Liu und Weihong Tan. „Aptamers generated by Cell SELEX for biomarker discovery". In: *Biomarkers in Medicine* 3.2 (Apr. 2009), S. 193–202. DOI: 10.2217/bmm.09.5 (siehe S. 48).

[143] Kevin N. Morris, Kirk B. Jensen, Carol M. Julin, Michael Weil und Larry Gold. „High affinity ligands from in vitro selection: Complex targets". In: *Proceedings of the National Academy of Sciences* 95.6 (1998), S. 2902–2907 (siehe S. 48).

[144] R White. „Generation of Species Cross-reactive Aptamers Using Toggle SELEX". In: *Molecular Therapy* 4.6 (Dez. 2001), S. 567–573. DOI: 10.1006/mthe.2001.0495 (siehe S. 48).

[145] L R Coulter, M A Landree und T A Cooper. „Identification of a new class of exonic splicing enhancers by in vivo selection." In: *Molecular and Cellular Biology* 17.4 (Apr. 1997), S. 2143–2150. DOI: 10.1128/mcb.17.4.2143 (siehe S. 48).

[146] D. H. Burke und J. H. Willis. „Recombination, RNA evolution, and bifunctional RNA molecules isolated through chimeric SELEX". In: *RNA* 4.9 (Sep. 1998), S. 1165–1175 (siehe S. 48).

[147] L Wu. „An allosteric synthetic DNA". In: *Nucleic Acids Research* 27.6 (März 1999), S. 1512–1516. DOI: 10.1093/nar/27.6.1512 (siehe S. 48).

[148] Guillermo Aquino-Jarquin und Julia D. Toscano-Garibay. „RNA Aptamer Evolution: Two Decades of SELEction“. In: *International Journal of Molecular Sciences* 12.12 (Dez. 2011), S. 9155–9171. DOI: 10.3390/ijms12129155 (siehe S. 48 f.).

[149] Anthony D Keefe und Sharon T Cload. „SELEX with modified nucleotides“. In: *Current Opinion in Chemical Biology* 12.4 (Aug. 2008), S. 448–456. DOI: 10.1016/j.cbpa.2008.06.028 (siehe S. 48).

[150] Razvan Nutiu und Yingfu Li. „In Vitro Selection of Structure-Switching Signaling Aptamers“. In: *Angewandte Chemie International Edition* 44.7 (Feb. 2005), S. 1061–1065. DOI: 10.1002/anie.200461848 (siehe S. 48).

[151] J. H. Davis und J. W. Szostak. „Isolation of high-affinity GTP aptamers from partially structured RNA libraries“. In: *Proceedings of the National Academy of Sciences* 99.18 (Aug. 2002), S. 11616–11621. DOI: 10.1073/pnas.182095699 (siehe S. 48).

[152] Karen M. Ruff, Thomas M. Snyder und David R. Liu. „Enhanced Functional Potential of Nucleic Acid Aptamer Libraries Patterned to Increase Secondary Structure“. In: *J. Am. Chem. Soc.* 132.27 (Juli 2010), S. 9453–9464. DOI: 10.1021/ja103023m (siehe S. 48).

[153] Drew Smith, Gary P. Kirschenheuter, Josephine Charlton, David M. Guidot und John E. Repine. „In vitro selection of RNA-based irreversible inhibitors of human neutrophil elastase“. In: *Chemistry & Biology* 2.11 (Nov. 1995), S. 741–750. DOI: 10.1016/1074-5521(95)90102-7 (siehe S. 48).

[154] K. B. Jensen, B. L. Atkinson, M. C. Willis, T. H. Koch und L. Gold. „Using in vitro selection to direct the covalent attachment of human immunodeficiency virus type 1 Rev protein to high-affinity RNA ligands“. In: *Proc. Natl. Acad. Sci. U.S.A.* 92.26 (Dez. 1995), S. 12220–12224 (siehe S. 48).

[155] Tomohiro Shimada, Nobuyuki Fujita, Michihisa Maeda und Akira Ishihama. „Systematic search for the Cra-binding promoters using genomic SELEX system“. In: *Genes to Cells* 10.9 (Sep. 2005), S. 907–918. DOI: 10.1111/j.1365-2443.2005.00888.x (siehe S. 49).

[156] A. Vater. „Short bioactive Spiegelmers to migraine-associated calcitonin gene-related peptide rapidly identified by a novel approach: Tailored-SELEX“. In: *Nucleic Acids Research* 31.21 (Nov. 2003), 130e–130. DOI: 10.1093/nar/gng130 (siehe S. 49, 159).

[157] J.-D. Wen. „Selection of genomic sequences that bind tightly to Ff gene 5 protein: primer-free genomic SELEX“. In: *Nucleic Acids Research* 32.22 (Dez. 2004), e182–e182. DOI: 10.1093/nar/gnh179 (siehe S. 49).

[158] Weihua Pan, Ping Xin, Susan Patrick *et al.* „Primer-Free Aptamer Selection Using A Random DNA Library“. In: *Journal of Visualized Experiments* 41 (Juli 2010). DOI: 10.3791/2039 (siehe S. 49, 159).

[159] Maxim Berezovski, Michael Musheev, Andrei Drabovich und Sergey N. Krylov. „Non-SELEX Selection of Aptamers“. In: *J. Am. Chem. Soc.* 128.5 (Feb. 2006), S. 1410–1411. DOI: 10.1021/ja056943j (siehe S. 49).

[160] A. Jolma, T. Kivioja, J. Toivonen *et al.* „Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities“. In: *Genome Research* 20.6 (Apr. 2010), S. 861–873. DOI: 10.1101/gr.100552.109 (siehe S. 49).

[161] D. Eulberg. „Development of an automated in vitro selection protocol to obtain RNA-based aptamers: identification of a biostable substance P antagonist“. In: *Nucleic Acids Research* 33.4 (Feb. 2005), e45–e45. DOI: 10.1093/nar/gni044 (siehe S. 49).

[162] Guizhao Liang und Zhiliang Li. „Scores of generalized base properties for quantitative sequence-activity modelings for E. coli promoters based on support vector machine“. In: *Journal of Molecular Graphics and Modelling* 26.1 (2007), S. 269–281 (siehe S. 52 f., 55, 61 f., 70).

[163] Ulf Norinder und Jörgen Jonsson. „Theoretical descriptors of nucleic acid bases. Application to DNA promoter sequences“. In: *Quantitative Structure-Activity Relationships* 13.3 (1994), S. 295–301 (siehe S. 52–55).

[164] Maria Sandberg, Michael Sjöström und Jörgen Jonsson. „A multivariate characterization of tRNA nucleosides“. In: *Journal of Chemometrics* 10.5-6 (1996), S. 493–508 (siehe S. 52 f., 61).

[165] P Broto, G Moreau und C Vandycke. „Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients“. In: *European journal of medicinal chemistry* 19.1 (1984), S. 71–78 (siehe S. 52, 56).

[166] Markus Wagener, Jens Sadowski und Johann Gasteiger. „Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks“. In: *J. Am. Chem. Soc.* 117.29 (Juli 1995), S. 7769–7775. DOI: 10.1021/ja00134a023 (siehe S. 52, 56).

[167] Patrick AP Moran. „Notes on continuous stochastic phenomena“. In: *Biometrika* (1950), S. 17–23 (siehe S. 52, 56).

[168] Robert C Geary. „The contiguity ratio and statistical mapping“. In: *The incorporated statistician* (1954), S. 115–146 (siehe S. 52, 56).

[169] Y. Marrero-Ponce. „Linear Indices of the Molecular Pseudographs Atom Adjacency Matrix: Definition, Significance-Interpretation, and Application to QSAR Analysis of Flavone Derivatives as HIV-1 Integrase Inhibitors“. In: *Journal of Chemical Information and Modeling* 44.6 (Nov. 2004), S. 2010–2026. DOI: 10.1021/ci049950k (siehe S. 52, 58).

[170] Yovani Marrero-Ponce, Francisco Torrens, Ramón García-Domenech, Sadiel E. Ortega-Broche und Vicente Romero Zaldivar. „Novel 2D TOMOCOMD-CARDD molecular descriptors: atom-based stochastic and non-stochastic bilinear indices and their QSPR applications“. In: *Journal of Mathematical Chemistry* 44.3 (Juni 2008), S. 650–673. DOI: 10.1007/s10910-008-9389-0 (siehe S. 52, 58).

[171] Andrea Mauri, Viviana Consonni, Manuela Pavan und Roberto Todeschini. „Dragon software: An easy approach to molecular descriptor calculations“. In: *MATCH Commun Math Comput Chem* 56 (2006), S. 237–248 (siehe S. 52 f., 59 f.).

[172] R Todeschini, V Consonni, A Mauri und M Pavan. *DRAGON version 6, Talete srl, Milan, Italy*. 2011 (siehe S. 52).

[173] M. Ganapathiraju, Madhavi Ganapathiraju, Vijayalaxmi Manoharan und Judith Klein-seetharaman. „BLMT - Statistical Sequence Analysis Using N-Grams“. In: *J. Applied Bioinformatics* 3 (2004) (siehe S. 52, 58).

[174] Roberto Todeschini und Viviana Consonni. *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)*. Bd. 41. John Wiley & Sons, 2009 (siehe S. 53, 56, 58).

[175] Peter Ertl, Bernhard Rohde und Paul Selzer. „Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties". In: *Journal of Medicinal Chemistry* 43.20 (Okt. 2000), S. 3714–3717. DOI: 10.1021/jm000942e (siehe S. 53).

[176] S. Liu, C. Cao und Z. Li. „Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector". In: *Journal of Chemical Information and Modeling* 38.3 (Mai 1998), S. 387–394. DOI: 10.1021/ci970109z (siehe S. 53).

[177] Shushen Liu, Chunsheng Yin, Shaoxi Cai und Zhiliang Li. „A Novel MHDV Descriptor for Dipeptide QSAR Studies". In: *Jnl Chinese Chemical Soc* 48.2 (Apr. 2001), S. 253–260. DOI: 10.1002/jccs.200100041 (siehe S. 53).

[178] J. Galvez, R. Garcia, M. T. Salabert und R. Soler. „Charge Indexes. New Topological Descriptors". In: *Journal of Chemical Information and Modeling* 34.3 (Mai 1994), S. 520–525. DOI: 10.1021/ci00019a008 (siehe S. 53).

[179] Alexandru T. Balaban, Dan Ciubotariu und Mihai Medeleanu. „Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors". In: *Journal of Chemical Information and Modeling* 31.4 (Nov. 1991), S. 517–523. DOI: 10.1021/ci00004a014 (siehe S. 53).

[180] Christoph Ruecker und Gerta Ruecker. „Mathematical Relation between Extended Connectivity and Eigenvector Coefficients". In: *Journal of Chemical Information and Modeling* 34.3 (Mai 1994), S. 534–538. DOI: 10.1021/ci00019a010 (siehe S. 53).

[181] Gerta Ruecker und Christoph Ruecker. „Counts of all walks as atomic and molecular descriptors". In: *Journal of Chemical Information and Modeling* 33.5 (Sep. 1993), S. 683–695. DOI: 10.1021/ci00015a005 (siehe S. 53).

[182] V. Consonni, R. Todeschini und M. Pavan. „Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors". In: *Journal of Chemical Information and Modeling* 42.3 (Mai 2002), S. 682–692. DOI: 10.1021/ci015504a (siehe S. 53).

[183] Roberto Todeschini und Paola Gramatica. „SD-modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors". In: *Quantitative Structure-Activity Relationships* 16.2 (1997), S. 113–119. DOI: 10.1002/qsar.19970160203 (siehe S. 53).

[184] Johann Gasteiger, Jens Sadowski, Jan Schuur *et al.* „Chemical Information in 3D Space". In: *Journal of Chemical Information and Computer Sciences* 36.5 (1996), S. 1030–1037. DOI: 10.1021/ci960343+ (siehe S. 53).

[185] J.H. Schuur, P. Selzer und J. Gasteiger. „The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity". In: *Journal of Chemical Information and Modeling* 36.2 (März 1996), S. 334–344. DOI: 10.1021/ci950164c (siehe S. 53).

[186] Milan Randic, Alexander F. Kleiner und Luz M. De Alba. „Distance/Distance Matrixes". In: *Journal of Chemical Information and Modeling* 34.2 (März 1994), S. 277–286. DOI: 10.1021/ci00018a008 (siehe S. 53).

[187] A.T. Balaban. „From Chemical Topology to 3D Geometry“. In: *Journal of Chemical Information and Modeling* 37.4 (Juli 1997), S. 645–650. DOI: 10.1021/ci960168x (siehe S. 53).

[188] Mircea V. Diudea, Dragos Horvath und Ante Graovac. „Molecular Topology. 15. 3D Distance Matrixes and Related Topological Indices“. In: *Journal of Chemical Information and Modeling* 35.1 (Jan. 1995), S. 129–135. DOI: 10.1021/ci00023a019 (siehe S. 53).

[189] Shushen Liu, Chunsheng Yin, Shaoxi Cai und Zhiliang Li. „A novel MHDV descriptor for dipeptide QSAR studies“. In: *Journal of the Chinese Chemical Society* 48.2 (2001), S. 253–260 (siehe S. 53).

[190] Ulf Norinder. „A theoretical reinvestigation of the nucleic bases adenine, guanine, cytosine, thymine and uracil using AM1“. In: *Journal of Molecular Structure: THEOCHEM* 151 (Mai 1987), S. 259–269. DOI: 10.1016/0166-1280(87)85062-5 (siehe S. 54).

[191] Prof.K. Steliou. *Model version 2.96*. Version 2.96. Department of Chemistry, University of Montreal (siehe S. 54).

[192] James J Stewart. *MOPAC manual. A general molecular orbital package*. Techn. Ber. DTIC Document, 1990 (siehe S. 54).

[193] Michael J. S. Dewar, Eve G. Zoebisch, Eamonn F. Healy und James J. P. Stewart. „Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model“. In: *J. Am. Chem. Soc.* 107.13 (Juni 1985), S. 3902–3909. DOI: 10.1021/ja00299a024 (siehe S. 54).

[194] Borries Demeler und Guangwen Zhou. „Neural network optimization for E.coli promoter prediction“. In: *Nucl Acids Res* 19.7 (1991), S. 1593–1599. DOI: 10.1093/nar/19.7.1593 (siehe S. 54, 61).

[195] Jögen Jonsson, Torbjörn Norberg, Lena Carlsson, Claes Gustafsson und Svante Wold. „Quantitative sequence-activity models (QSAM)—tools for sequence design“. In: *Nucl Acids Res* 21.3 (1993), S. 733–739. DOI: 10.1093/nar/21.3.733 (siehe S. 54, 61 f.).

[196] A. Bogan-Marta, A. Hategan und I. Pitas. „Language engineering and information theoretic methods in protein sequence similarity studies“. In: *Computational Intelligence in Medical Informatics*. Springer Science + Business Media, 2008, S. 151–183. DOI: 10.1007/978-3-540-75767-2_8 (siehe S. 58 f.).

[197] S. C. Flores und R. B. Altman. „Turning limited experimental information into 3D models of RNA“. In: *RNA* 16.9 (Juli 2010), S. 1769–1778. DOI: 10.1261/rna.2112110 (siehe S. 59, 185).

[198] Andrew V. Colasanti, Xiang-Jun Lu und Wilma K. Olson. „Analyzing and Building Nucleic Acid Structures with 3DNA“. In: *Journal of Visualized Experiments* 74 (2013). DOI: 10.3791/4401 (siehe S. 59, 62, 186).

[199] Alan R Katritzky, Ruslan Petrukhin, Hongfang Yang und Mati Karelson. „Codessa Pro“. In: *User's manual. University of Florida* (2002) (siehe S. 60).

[200] Molecular Networks GmbH. „CORINA Symphony“. In: *User's manual. Molecular Networks.* (2014) (siehe S. 60).

[201] *Molecular Operating Environment (MOE)*. Version 2013.08. 1010 Sherbooke St. West, Suite Nr.910, Montreal, QC, Canada, H3A 2R7: Chemical Computing Group Inc., 2015 (siehe S. 60).

[202] Christoph Steinbeck, Christian Hoppe, Stefan Kuhn *et al.* „Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics“. In: *Current pharmaceutical design* 12.17 (2006), S. 2111–2120 (siehe S. 60).

[203] Chun Wei Yap. „PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints“. In: *Journal of computational chemistry* 32.7 (2011), S. 1466–1474 (siehe S. 59 f., 63, 71).

[204] Ángel Durán, Guillermo C. Martínez und Manuel Pastor. „Development and Validation of AMANDA, a New Algorithm for Selecting Highly Relevant Regions in Molecular Interaction Fields“. In: *Journal of Chemical Information and Modeling* 48.9 (Sep. 2008), S. 1813–1823. DOI: 10.1021/ci800037t (siehe S. 60).

[205] G. Cruciani, P. Crivori, P.-A. Carrupt und B. Testa. „Molecular fields in quantitative structure–permeation relationships: the VolSurf approach“. In: *Journal of Molecular Structure: THEOCHEM* 503.1-2 (Mai 2000), S. 17–30. DOI: 10.1016/s0166-1280(99)00360-7 (siehe S. 60).

[206] „ISIDA Fragmentor2015“. In: *User manual* (2015) (siehe S. 60).

[207] *AFGen: Fragment-based Descriptors for Chemical Compounds*. Version 2.0 (siehe S. 60).

[208] Liying Zhang, Denis Fourches, Alexander Sedykh *et al.* „Discovery of Novel Antimalarial Compounds Enabled by QSAR-Based Virtual Screening“. In: *Journal of Chemical Information and Modeling* 53.2 (Feb. 2013), S. 475–492. DOI: 10.1021/ci300421n (siehe S. 59).

[209] Paola Gramatica, Stefano Cassani und Nicola Chirico. „QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS“. In: *J. Comput. Chem.* 35.13 (März 2014), S. 1036–1044. DOI: 10.1002/jcc.23576 (siehe S. 59).

[210] Douglas F. Browning und Stephen J. W. Busby. „The regulation of bacterial transcription initiation“. In: *Nat Rev Micro* 2.1 (Jan. 2004), S. 57–65. DOI: 10.1038/nrmicro787 (siehe S. 61).

[211] Martin Jinek und Jennifer A. Doudna. „A three-dimensional view of the molecular machinery of RNA interference“. In: *Nature* 457.7228 (Jan. 2009), S. 405–412. DOI: 10.1038/nature07755 (siehe S. 61).

[212] Eva Yus, Marc Güell, Ana P Vivancos *et al.* „Transcription start site associated RNAs in bacteria“. In: *Mol Syst Biol* 8 (Mai 2012). DOI: 10.1038/msb.2012.16 (siehe S. 61, 75).

[213] M. Rosenberg und D. Court. „Regulatory sequences involved in the promotion and termination of RNA transcription“. In: *Annu. Rev. Genet.* 13 (1979), S. 319–353 (siehe S. 61, 75).

[214] D. Pribnow. „Bacteriophage T7 early promoters: nucleotide sequences of two RNA polymerase binding sites“. In: *J. Mol. Biol.* 99.3 (Dez. 1975), S. 419–443 (siehe S. 61, 75).

[215] Martin E. Mulligan, Diane K. Hawley, Robert Entriken und William R. McClure. „Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity“. In: *Nucl Acids Res* 12.1Part2 (1984), S. 789–800. DOI: 10.1093/nar/12.1part2.789 (siehe S. 61).

[216] H. Kiryu, T. Oshima und K. Asai. „Extracting relations between promoter sequences and their strengths from microarray data“. In: *Bioinformatics* 21.7 (Okt. 2004), S. 1062–1068. DOI: 10.1093/bioinformatics/bti094 (siehe S. 61, 77).

[217] Makoto Kobayashi, Kyosuke Nagata und Akira Ishihama. „Promoter selectivity of Escherichia coli RNA polymerase: effect of base substitutions in the promoter -35 region on promoter strength“. In: *Nucl Acids Res* 18.24 (1990), S. 7367–7372. DOI: 10.1093/nar/18.24.7367 (siehe S. 61).

[218] Deborah G. Ayers, David T. Auble und Pieter L. deHaseth. „Promoter recognition by Escherichia coli RNA polymerase“. In: *Journal of Molecular Biology* 207.4 (Juni 1989), S. 749–756. DOI: 10.1016/0022-2836(89)90241-6 (siehe S. 61).

[219] M. Lanzer und H. Bujard. „Promoters largely determine the efficiency of repressor action“. In: *Proc. Natl. Acad. Sci. U.S.A.* 85.23 (Dez. 1988), S. 8973–8977 (siehe S. 61).

[220] U. Deuschle, W. Kammerer, R. Gentz und H. Bujard. „Promoters of Escherichia coli: a hierarchy of in vivo strength indicates alternate structures“. In: *EMBO J.* 5.11 (Nov. 1986), S. 2987–2994 (siehe S. 61).

[221] T. Yamagishi, S. Hirose und T. Kondo. „Secondary DNA structure formation for Hoxb9 promoter and identification of its specific binding protein“. In: *Nucleic Acids Research* 36.6 (Feb. 2008), S. 1965–1975. DOI: 10.1093/nar/gkm1079 (siehe S. 62).

[222] Bulukani Mlalazi. „Defining the role of phytoene synthase in carotenoid accumulation of high provitamin A bananas“. Diss. Queensland University of Technology, 2010 (siehe S. 62).

[223] Ronny Lorenz, Stephan HF Bernhart, Christian Hoener Zu Siederdissen *et al.* „ViennaRNA Package 2.0.“ In: *Algorithms for Molecular Biology* 6.1 (2011), S. 26 (siehe S. 62).

[224] Cort J Willmott und Kenji Matsuura. „Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance“. In: *Climate research* 30.1 (2005), S. 79 (siehe S. 65).

[225] Cort J. Willmott, Kenji Matsuura und Scott M. Robeson. „Ambiguities inherent in sums-of-squares-based error statistics“. In: *Atmospheric Environment* 43.3 (Jan. 2009), S. 749–752. DOI: 10.1016/j.atmosenv.2008.10.005 (siehe S. 65).

[226] T. Chai und R. R. Draxler. „Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature“. In: *Geosci. Model Dev.* 7.3 (2014), S. 1247–1250. DOI: 10.5194/gmd-7-1247-2014 (siehe S. 65).

[227] Feng-Jenq Lin. „Solving Multicollinearity in the Process of Fitting Regression Model Using the Nested Estimate Procedure“. In: *Quality & Quantity* 42.3 (Dez. 2006), S. 417–426. DOI: 10.1007/s11135-006-9055-1 (siehe S. 66).

[228] John Neter, William Wasserman und Michael H Kutner. „Applied linear regression models“. In: (1989) (siehe S. 66).

[229] A.-L. Boulesteix und K. Strimmer. „Partial least squares: a versatile tool for the analysis of high-dimensional genomic data“. In: *Briefings in Bioinformatics* 8.1 (Mai 2006), S. 32–44. DOI: 10.1093/bib/bbl016 (siehe S. 66).

[230] Max Kuhn und Kjell Johnson. *Applied predictive modeling*. Springer Science + Business Media, 2013. DOI: 10.1007/978-1-4614-6849-3 (siehe S. 66).

[231] Trygve Almøy. „A simulation study on comparison of prediction methods when only a few components are relevant“. In: *Computational Statistics & Data Analysis* 21.1 (Jan. 1996), S. 87–107. DOI: 10.1016/0167-9473(95)00006-2 (siehe S. 66).

[232] lldiko E. Frank und Jerome H. Friedman. „A Statistical View of Some Chemometrics Regression Tools“. In: *Technometrics* 35.2 (Mai 1993), S. 109–135. DOI: 10.1080/00401706.1993.10485033 (siehe S. 66).

[233] D. V. Nguyen und D. M. Rocke. „Tumor classification by partial least squares using microarray gene expression data“. In: *Bioinformatics* 18.1 (Jan. 2002), S. 39–50. DOI: 10.1093/bioinformatics/18.1.39 (siehe S. 66).

[234] Jonas Nilsson, Sijmen de Jong und Age K. Smilde. „Multiway calibration in 3D QSAR“. In: *Journal of Chemometrics* 11.6 (1997), S. 511–524. ISSN: 1099-128X. DOI: 10.1002/(SICI)1099-128X(199711/12)11:6<511::AID-CEM488>3.0.CO;2-W (siehe S. 66, 79).

[235] Humberto González-Díaz, Alcides Pérez-Bello, Eugenio Uriarte und Yenny González-Díaz. „QSAR study for mycobacterial promoters with low sequence homology“. In: *Bioorganic & Medicinal Chemistry Letters* 16.3 (Feb. 2006), S. 547–553. DOI: 10.1016/j.bmcl.2005.10.057 (siehe S. 66).

[236] M.J. Sorich, J.O. Miners, R.A. McKinnon *et al.* „Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms“. In: *Journal of Chemical Information and Modeling* 43.6 (Nov. 2003), S. 2019–2024. DOI: 10.1021/ci034108k (siehe S. 66).

[237] Farhad Gharagheizi. „QSPR Studies for Solubility Parameter by Means of Genetic Algorithm–Based Multivariate Linear Regression and Generalized Regression Neural Network“. In: *QSAR & Combinatorial Science* 27.2 (Feb. 2008), S. 165–170. DOI: 10.1002/qsar.200630159 (siehe S. 66).

[238] Roman Rosipal und Nicole Krämer. „Overview and recent advances in partial least squares“. In: *Subspace, latent structure and feature selection*. Springer, 2006, S. 34–51 (siehe S. 66).

[239] S. Wold, H. Martens und H. Wold. „The multivariate calibration problem in chemistry solved by the PLS method“. In: *Matrix Pencils*. Springer Science + Business Media, 1982, S. 286–293. DOI: 10.1007/bfb0062108 (siehe S. 66).

[240] Hannes Feilhauer, Gregory P. Asner, Roberta E. Martin und Sebastian Schmidtlein. „Brightness–normalized Partial Least Squares Regression for hyperspectral data“. In: *Journal of Quantitative Spectroscopy and Radiative Transfer* 111.12-13 (Aug. 2010), S. 1947–1957. DOI: 10.1016/j.jqsrt.2010.03.007 (siehe S. 66).

[241] E. Frank, M. Hall, L. Trigg, G. Holmes und I. H. Witten. „Data mining in bioinformatics using Weka“. In: *Bioinformatics* 20.15 (Apr. 2004), S. 2479–2481. DOI: 10.1093/bioinformatics/bth261 (siehe S. 67).

[242] Mark Hall, Eibe Frank, Geoffrey Holmes *et al.* „The WEKA data mining software“. In: *ACM SIGKDD Explorations Newsletter* 11.1 (Nov. 2009), S. 10. DOI: 10.1145/1656274.1656278 (siehe S. 67).

[243] Tahir Mehmood, Kristian Hovde Liland, Lars Snipen und Solve Sæbø. „A review of variable selection methods in partial least squares regression“. In: *Chemometrics and Intelligent Laboratory Systems* 118 (2012), S. 62–69 (siehe S. 67).

[244] Riccardo Leardi und Amparo Lupiáñez González. „Genetic algorithms applied to feature selection in PLS regression: how and when to use them“. In: *Chemometrics and Intelligent Laboratory Systems* 41.2 (1998), S. 195–207 (siehe S. 67).

[245] Jahanbakhsh Ghasemi und Shahin Ahmadi. „Combination of genetic algorithm and partial least squares for cloud point prediction of nonionic surfactants from molecular structures“. In: *Annali di chimica* 97.1-2 (2007), S. 69–83 (siehe S. 67).

[246] B. L. WELCH. „THE GENERALIZATION OF STUDENTS PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARIANCES ARE INVOLVED“. In: *Biometrika* 34.1-2 (1947), S. 28–35. DOI: 10.1093/biomet/34.1-2.28 (siehe S. 67).

[247] STUDENT. „THE PROBABLE ERROR OF A MEAN“. In: *Biometrika* 6.1 (März 1908), S. 1–25. DOI: 10.1093/biomet/6.1.1 (siehe S. 67).

[248] H. B. Mann und D. R. Whitney. „On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other“. In: *Ann. Math. Statist.* 18.1 (März 1947), S. 50–60. DOI: 10.1214/aoms/1177730491 (siehe S. 68).

[249] Eva Skovlund und Grete U. Fenstad. „Should we always choose a nonparametric test when comparing two apparently nonnormal distributions?“ In: *Journal of Clinical Epidemiology* 54.1 (Jan. 2001), S. 86–92. DOI: 10.1016/s0895-4356(00)00264-x (siehe S. 68).

[250] Morten W Fagerland. „t-tests, non-parametric tests, and large studies—a paradox of statistical practice?“ In: *BMC Medical Research Methodology* 12.1 (2012), S. 78. DOI: 10.1186/1471-2288-12-78 (siehe S. 68).

[251] Morten W. Fagerland und Leiv Sandvik. „The Wilcoxon-Mann-Whitney test under scrutiny“. In: *Statist. Med.* 28.10 (Mai 2009), S. 1487–1497. DOI: 10.1002/sim.3561 (siehe S. 68).

[252] Rani T. Sobha und Bapi Raju S. „Analysis of n-Gram based Promoter Recognition Methods and Application to Whole Genome Promoter Prediction“. In: *In Silico Biology* 9.1, 2 (2009), S1–S16. ISSN: 1386-6338. DOI: 10.3233/ISB-2009-0388 (siehe S. 73).

[253] Chandrashekar Jatoth und Rupesh Mahajan. „AnG-HPR: Analysis of n-Gram based human Promoter Recognition“. In: *International Journal of Engineering Research and Applications* 2 (2012), S. 247–254. ISSN: 2248-9622 (siehe S. 73).

[254] W. Gruissem und G. Zurawski. „Identification and mutational analysis of the promoter for a spinach chloroplast transfer RNA gene“. In: *EMBO J.* 4.7 (Juli 1985), S. 1637–1644 (siehe S. 77).

[255] T. J. Kenney und G. Churchward. „Genetic analysis of the Mycobacterium smegmatis rpsL promoter“. In: *J. Bacteriol.* 178.12 (Juni 1996), S. 3564–3571 (siehe S. 77).

[256] Sreenivasan Ponnambalam, Christine Webster, Alistair Bingham und Stephen Busby. „Transcription initiation at the Escherichia coli galactose operon promoters in the absence of the normal-35 region sequences.“ In: *Journal of Biological Chemistry* 261.34 (1986), S. 16043–16048 (siehe S. 77).

[257] S. S. Singh, A. Typas, R. Hengge und D. C. Grainger. „Escherichia coli 70 senses sequence and conformation of the promoter spacer region“. In: *Nucleic Acids Research* 39.12 (März 2011), S. 5109–5118. DOI: 10.1093/nar/gkr080 (siehe S. 78).

[258] Marianne Defernez und E.Katherine Kemsley. „The use and misuse of chemometrics for treating classification problems“. In: *TrAC Trends in Analytical Chemistry* 16.4 (Apr. 1997), S. 216–221. DOI: 10.1016/s0165-9936(97)00015-0 (siehe S. 79).

[259] D.M. Hawkins. „The Problem of Overfitting“. In: *Journal of Chemical Information and Modeling* 44.1 (Jan. 2004), S. 1–12. DOI: 10.1021/ci0342472 (siehe S. 79).

[260] Jun Shao. „Linear Model Selection by Cross-validation“. In: *Journal of the American Statistical Association* 88.422 (Juni 1993), S. 486–494. DOI: 10.1080/01621459.1993.10476299 (siehe S. 79).

[261] Knut Baumann. „Cross-validation as the objective function for variable-selection techniques“. In: *TrAC Trends in Analytical Chemistry* 22.6 (Juni 2003), S. 395–406. DOI: 10.1016/s0165-9936(03)00607-1 (siehe S. 79).

[262] Riccardo Leardi. „Application of genetic algorithm-PLS for feature selection in spectral data sets“. In: *J. Chemometrics* 14.5-6 (2000), S. 643–655. DOI: 10.1002/1099-128x(200009/12)14:5/6<643::aid-cem621>3.0.co;2-e (siehe S. 79).

[263] D. Jouan-Rimbaud, D.L. Massart und O.E. de Noord. „Random correlation in variable selection for multivariate calibration with a genetic algorithm“. In: *Chemometrics and Intelligent Laboratory Systems* 35.2 (Dez. 1996), S. 213–220. DOI: 10.1016/s0169-7439(96)00062-7 (siehe S. 79).

[264] Fredrik Lindgren, Björn Hansen, Walter Karcher, Michael Sjöström und Lennart Eriksson. „Model validation by permutation tests: Applications to variable selection“. In: *J. Chemometrics* 10.5-6 (Sep. 1996), S. 521–532. DOI: 10.1002/(sici)1099-128x(199609)10:5/6<521::aid-cem448>3.0.co;2-j (siehe S. 79).

[265] Riccardo Leardi und Amparo Lupiáñez González. „Genetic algorithms applied to feature selection in PLS regression: how and when to use them“. In: *Chemometrics and Intelligent Laboratory Systems* 41.2 (Juli 1998), S. 195–207. DOI: 10.1016/s0169-7439(98)00051-3 (siehe S. 79).

[266] Christoph Rücker, Gerta Rücker und Markus Meringer. „y-Randomization and Its Variants in QSPR/QSAR“. In: *Journal of Chemical Information and Modeling* 47.6 (Nov. 2007), S. 2345–2357. DOI: 10.1021/ci700157b (siehe S. 79).

[267] Gary D. Stormo, Thomas D. Schneider, Larry Gold und Andrzej Ehrenfeucht. „Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in E. coli“. In: *Nucl Acids Res* 10.9 (1982), S. 2997–3011. DOI: 10.1093/nar/10.9.2997 (siehe S. 84).

[268] C. E. Shannon. „A Mathematical Theory of Communication“. In: *Bell System Technical Journal* 27.4 (Okt. 1948), S. 623–656. DOI: 10.1002/j.1538-7305.1948.tb00917.x (siehe S. 85).

[269] John C. Wootton und Scott Federhen. „[33] Analysis of compositionally biased regions in sequence databases“. In: *Methods in Enzymology*. Elsevier BV, 1996, S. 554–571. DOI: 10.1016/s0076-6879(96)66035-2 (siehe S. 86).

[270] Rainer Merkl und Stephan Waack. *Bioinformatik Interaktiv - Grundlagen, Algorithmen, Anwendungen*. New York: John Wiley & Sons, 2013. ISBN: 978-3-527-68274-4 (siehe S. 86).

[271] Leonhard Euler. „De progressionibus transcendentibus seu quarum termini generales algebraice dari nequeunt“. Lateinisch. In: *Commentarii academiae scientiarum Petropolitanae* 5 (1738), S. 36–57 (siehe S. 86).

[272] Adrien-Marie Legendre. „Recherches sur diverses sortes d’intégrales définies“. Französisch. In: *Mémoires de la classe des sciences mathématiques et physiques de l’Institut de France* 10 (13. Nov. 1809), S. 477 (siehe S. 86).

[273] Thomas D. Schneider und R.Michael Stephens. „Sequence logos: a new way to display consensus sequences“. In: *Nucl Acids Res* 18.20 (1990), S. 6097–6100. DOI: 10.1093/nar/18.20.6097 (siehe S. 87).

[274] L. KARTTUNEN, J-P. CHANOD, G. GREFENSTETTE und A. SCHILLE. „Regular expressions for language engineering“. In: *Natural Language Engineering* 2.4 (Dez. 1996), S. 305–328. DOI: 10.1017/s1351324997001563 (siehe S. 88).

[275] David Maier. „The Complexity of Some Problems on Subsequences and Supersequences“. In: *Journal of the ACM* 25.2 (Apr. 1978), S. 322–336. DOI: 10.1145/322063.322075 (siehe S. 88).

[276] F. Sievers, A. Wilm, D. Dineen *et al.* „Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega“. In: *Molecular Systems Biology* 7.1 (Apr. 2011), S. 539–539. DOI: 10.1038/msb.2011.75 (siehe S. 89).

[277] Itamar Sela, Haim Ashkenazy, Kazutaka Katoh und Tal Pupko. „GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters“. In: *Nucleic Acids Res* 43.W1 (Apr. 2015), W7–W14. DOI: 10.1093/nar/gkv318 (siehe S. 89).

[278] Julie D. Thompson, Benjamin Linard, Odile Lecompte und Olivier Poch. „A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives“. In: *PLoS ONE* 6.3 (März 2011). Hrsg. von Jonathan Badger, e18093. DOI: 10.1371/journal.pone.0018093 (siehe S. 89).

[279] Stefano Iantorno, Kevin Gori, Nick Goldman, Manuel Gil und Christophe Dessimoz. „Who Watches the Watchmen? An Appraisal of Benchmarks for Multiple Sequence Alignment“. In: *Methods in Molecular Biology*. Springer Science + Business Media, Aug. 2013, S. 59–73. DOI: 10.1007/978-1-62703-646-7_4 (siehe S. 89).

[280] E. Ukkonen. „On-line construction of suffix trees“. In: *Algorithmica* 14.3 (Sep. 1995), S. 249–260. DOI: 10.1007/bf01206331 (siehe S. 89).

[281] Peter Weiner. „Linear pattern matching algorithms“. In: *14th Annual Symposium on Switching and Automata Theory (swat 1973)*. Institute of Electrical & Electronics Engineers (IEEE), Okt. 1973. DOI: 10.1109/swat.1973.13 (siehe S. 89 f.).

[282] Bieganski, Riedl, Cartis und Retzel. „Generalized suffix trees for biological sequence data: applications and implementation“. In: *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS-94*. Institute of Electrical & Electronics Engineers (IEEE), 1994. DOI: 10.1109/hicss.1994.323593 (siehe S. 90).

[283] Dan Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997. ISBN: 0521585198 (siehe S. 90).

[284] Faras Rasheed, Mohammed Alshalalfa und Reda Alhajj. „Efficient Periodicity Mining in Time Series Databases Using Suffix Trees“. In: *IEEE Trans. Knowl. Data Eng.* 23.1 (Jan. 2011), S. 79–94. DOI: 10.1109/tkde.2010.76 (siehe S. 90).

[285] Marina Barsky, Ulrike Stege, Alex Thomo und Chris Upton. „A new method for indexing genomes using on-disk suffix trees“. In: *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*. Association for Computing Machinery (ACM), 2008. DOI: 10.1145/1458082.1458170 (siehe S. 90).

[286] Ricardo A. Baeza-Yates und Gaston H. Gonnet. „Fast text searching for regular expressions or automaton searching on tries“. In: *Journal of the ACM* 43.6 (Nov. 1996), S. 915–936. DOI: 10.1145/235809.235810 (siehe S. 90).

[287] Hung Chim und Xiaotie Deng. „A new suffix tree similarity measure for document clustering“. In: *Proceedings of the 16th international conference on World Wide Web - WWW '07*. Association for Computing Machinery (ACM), 2007. DOI: 10.1145/1242572.1242590 (siehe S. 90).

[288] Paolo Ferragina, Raffaele Giancarlo, Giovanni Manzini und Marinella Sciortino. „Boosting textual compression in optimal linear time“. In: *Journal of the ACM* 52.4 (Juli 2005), S. 688–713. DOI: 10.1145/1082036.1082043 (siehe S. 90).

[289] A. Brazma, I. Jonassen, J. Vilo und E. Ukkonen. „Predicting gene regulatory elements in silico on a genomic scale“. In: *Genome Res.* 8.11 (Nov. 1998), S. 1202–1215 (siehe S. 92).

[290] R. J. Dakin. „A tree-search algorithm for mixed integer programming problems“. In: *The Computer Journal* 8.3 (März 1965), S. 250–255. DOI: 10.1093/comjnl/8.3.250 (siehe S. 95).

[291] Jean-Michel Claverie. „Some useful statistical properties of position-weight matrices“. In: *Computers & Chemistry* 18.3 (Sep. 1994), S. 287–294. DOI: 10.1016/0097-8485(94)85024-0 (siehe S. 99).

[292] Michael Levitt und Mark Gerstein. „A unified statistical framework for sequence comparison and structure comparison“. In: *Proceedings of the National Academy of Sciences* 95.11 (1998), S. 5913–5920 (siehe S. 99).

[293] William R Pearson. „Empirical statistical estimates for sequence similarity searches“. In: *Journal of Molecular Biology* 276.1 (Feb. 1998), S. 71–84. DOI: 10.1006/jmbi.1997.1525 (siehe S. 99).

[294] John A. Gerlt und Patricia C. Babbitt. „D IVERGENT E VOLUTION OF E NZYMATIC F UNCTION : Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies“. In: *Annu. Rev. Biochem.* 70.1 (Juni 2001), S. 209–246. DOI: 10.1146/annurev.biochem.70.1.209 (siehe S. 99).

[295] Eric D. Scheeff und Philip E. Bourne. „Structural Evolution of the Protein Kinase–Like Superfamily“. In: *PLoS Comp Biol* 1.5 (2005), e49. DOI: 10.1371/journal.pcbi.0010049 (siehe S. 99).

[296] L. Xie, L. Xie und P. E. Bourne. „A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery“. In: *Bioinformatics* 25.12 (Mai 2009), S. i305–i312. DOI: 10.1093/bioinformatics/btp220 (siehe S. 99).

[297] Walter A. Baase, Lijun Liu, Dale E. Tronrud und Brian W. Matthews. „Lessons from the lysozyme of phage T4“. In: *Protein Science* 19.4 (Jan. 2010), S. 631–641. DOI: 10.1002/pro.344 (siehe S. 99).

[298] Jooyoun Kang, Jiwon Jung und Seong Keun Kim. „Flexibility of single-stranded DNA measured by single-molecule FRET“. In: *Biophysical Chemistry* 195 (Dez. 2014), S. 49–52. DOI: 10.1016/j.bpc.2014.08.004 (siehe S. 99).

[299] J. W. Keepers, P. A. Kollman, P. K. Weiner und T. L. James. „Molecular mechanical studies of DNA flexibility: coupled backbone torsion angles and base-pair openings“. In: *Proc. Natl. Acad. Sci. U.S.A.* 79.18 (Sep. 1982), S. 5537–5541 (siehe S. 99).

[300] Lei Bao, Xi Zhang, Lei Jin und Zhi-Jie Tan. „Flexibility of nucleic acids: From DNA to RNA“. In: *Chinese Physics B* 25.1 (Jan. 2016), S. 018703. DOI: 10.1088/1674-1056/25/1/018703 (siehe S. 99).

[301] Ke Yu Wang, Sarah McCurdy, Regan G. Shea, S. Swaminathan und Philip H. Bolton. „A DNA aptamer which binds to and inhibits thrombin exhibits a new structural motif for DNA“. In: *Biochemistry* 32.8 (März 1993), S. 1899–1904. DOI: 10.1021/bi00059a003 (siehe S. 99).

[302] Angela S Whatley, Mark A Ditzler, Margaret J Lange *et al.* „Potent Inhibition of HIV-1 Reverse Transcriptase and Replication by Nonpseudoknot, “UCAA-motif” RNA Aptamers“. In: *Molecular Therapy — Nucleic Acids* 2.2 (Feb. 2013), e71. DOI: 10.1038/mtna.2012.62 (siehe S. 99).

[303] Y. Nomura, Y. Tanaka, J.-i. Fukunaga *et al.* „Solution structure of a DNA mimicking motif of an RNA aptamer against transcription factor AML1 Runt domain“. In: *Journal of Biochemistry* 154.6 (Aug. 2013), S. 513–519. DOI: 10.1093/jb/mvt082 (siehe S. 99).

[304] Eva Nogales. „The development of cryo-EM into a mainstream structural biology technique“. In: *Nature Methods* 13.1 (Dez. 2015), S. 24–27. DOI: 10.1038/nmeth.3694 (siehe S. 100).

[305] H. M. Berman. „The Protein Data Bank“. In: *Nucleic Acids Research* 28.1 (Jan. 2000), S. 235–242. DOI: 10.1093/nar/28.1.235 (siehe S. 100, 108, 118, 122).

[306] *PDB Current Holdings Breakdown.* 1. Feb. 2016. URL: http://www.rcsb.org/pdb/statistics/holdings.do (siehe S. 100).

[307] Shuchismita Dutta und Helen M. Berman. „Large Macromolecular Complexes in the Protein Data Bank: A Status Report“. In: *Structure* 13.3 (März 2005), S. 381–388. DOI: 10.1016/j.str.2005.01.008 (siehe S. 100).

[308] Alasdair C. Steven und Wolfgang Baumeister. „The future is hybrid“. In: *Journal of Structural Biology* 163.3 (Sep. 2008), S. 186–195. DOI: 10.1016/j.jsb.2008.06.002 (siehe S. 100).

[309] Marc C Deller und Bernhard Rupp. „Crystallisation of Proteins and Macromolecular Complexes: Past, Present and Future“. In: *eLS*. John Wiley & Sons, Ltd, 2014. ISBN: 9780470015902. DOI: 10.1002/9780470015902.a0002718.pub2 (siehe S. 100).

[310] Joseph R Luft, Edward H Snell und George T DeTitta. „Lessons from high-throughput protein crystallization screening: 10 years of practical experience“. In: *Expert Opinion on Drug Discovery* 6.5 (März 2011), S. 465–480. DOI: 10.1517/17460441.2011.566857 (siehe S. 100).

[311] Boundless. „Electric vs. Magnetic Forces“. In: *Boundless Physics* (26. Mai 2016) (siehe S. 101).

[312] Boundless. „Hydrogen Bonding“. In: *Boundless Chemistry* (26. Mai 2016) (siehe S. 101).

[313] Wikimedia Commons. *Strukturmodell eines DNA-Moleküls.* URL: https://commons.wikimedia.org/wiki/File:DNA_simple2.svg (besucht am 10. 08. 2016) (siehe S. 101).

[314] Wikimedia Commons. *Main protein structures levels*. URL: https://commons.wikimedia.org/wiki/File:Main_protein_structure_levels_en.svg (besucht am 10. 08. 2016) (siehe S. 101).

[315] Beisi Xu, Yuedong Yang, Haojun Liang und Yaoqi Zhou. „An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles“. In: *Proteins* 76.3 (Aug. 2009), S. 718–730. DOI: 10.1002/prot.22384 (siehe S. 100 f., 104 f., 117).

[316] William W. Chen und Eugene I. Shakhnovich. „Lessons from the design of a novel atomic potential for protein folding“. In: *Protein Sci.* 14.7 (Juli 2005), S. 1741–1752. DOI: 10.1110/ps.051440705 (siehe S. 100).

[317] I. A. Hubner, E. J. Deeds und E. I. Shakhnovich. „High-resolution protein folding with a transferable potential“. In: *Proceedings of the National Academy of Sciences* 102.52 (Dez. 2005), S. 18914–18919. DOI: 10.1073/pnas.0502181102 (siehe S. 100).

[318] J. E. Donald, W. W. Chen und E. I. Shakhnovich. „Energetics of protein-DNA interactions“. In: *Nucleic Acids Research* 35.4 (Jan. 2007), S. 1039–1047. DOI: 10.1093/nar/gkl1103 (siehe S. 100, 114, 117).

[319] S. Jones. „Protein-RNA interactions: a structural analysis“. In: *Nucleic Acids Research* 29.4 (Feb. 2001), S. 943–954. DOI: 10.1093/nar/29.4.943 (siehe S. 102).

[320] Michèle Treger und Eric Westhof. „Statistical analysis of atomic contacts at RNA-protein interfaces“. In: *J. Mol. Recognit.* 14.4 (2001), S. 199–214. DOI: 10.1002/jmr.534 (siehe S. 102).

[321] Diane Lejeune, Nicolas Delsaux, Benoît Charloteaux, Annick Thomas und Robert Brasseur. „Protein-nucleic acid recognition: Statistical analysis of atomic interactions and influence of DNA structure“. In: *Proteins* 61.2 (Aug. 2005), S. 258–271. DOI: 10.1002/prot.20607 (siehe S. 102).

[322] Hyunwoo Kim, Euna Jeong, Seong-Wook Lee und Kyungsook Han. „Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns“. In: *FEBS Letters* 552.2-3 (Aug. 2003), S. 231–239. DOI: 10.1016/s0014-5793(03)00930-x (siehe S. 102).

[323] E. Jeong, H. Kim, S. W. Lee und K. Han. „Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes“. In: *Mol. Cells* 16.2 (Okt. 2003), S. 161–167 (siehe S. 102).

[324] Jonathan J. Ellis, Mark Broom und Susan Jones. „Protein-RNA interactions: Structural analysis and functional classes“. In: *Proteins* 66.4 (Dez. 2006), S. 903–911. DOI: 10.1002/prot.21211 (siehe S. 102 f., 113).

[325] M. Gao und J. Skolnick. „DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions“. In: *Nucleic Acids Research* 36.12 (Mai 2008), S. 3978–3992. DOI: 10.1093/nar/gkn332 (siehe S. 102).

[326] O. T. P. Kim, K. Yura und N. Go. „Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction“. In: *Nucleic Acids Research* 34.22 (Nov. 2006), S. 6450–6460. DOI: 10.1093/nar/gkl819 (siehe S. 102).

[327] S. Jones. „Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins“. In: *Nucleic Acids Research* 31.24 (Dez. 2003), S. 7189–7198. DOI: 10.1093/nar/gkg922 (siehe S. 102).

[328] N. M. Luscombe, R. A. Laskowski und J. M. Thornton. „Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level“. In: *Nucleic Acids Res.* 29.13 (Juli 2001), S. 2860–2874 (siehe S. 102).

[329] N. Morozova, J. Allers, J. Myers und Y. Shamoo. „Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures“. In: *Bioinformatics* 22.22 (Sep. 2006), S. 2746–2752. DOI: 10.1093/bioinformatics/btl470 (siehe S. 102).

[330] R. P. Bahadur, M. Zacharias und J. Janin. „Dissecting protein-RNA recognition sites“. In: *Nucleic Acids Research* 36.8 (Feb. 2008), S. 2705–2716. DOI: 10.1093/nar/gkn102 (siehe S. 102).

[331] Z. Liu. „Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential“. In: *Nucleic Acids Research* 33.2 (Jan. 2005), S. 546–558. DOI: 10.1093/nar/gki204 (siehe S. 103, 113).

[332] Irina Tuszynska und Janusz M Bujnicki. „DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking“. In: *BMC Bioinformatics* 12.1 (2011), S. 348. DOI: 10.1186/1471-2105-12-348 (siehe S. 103 f.).

[333] Y. Chen. „A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys“. In: *Nucleic Acids Research* 32.17 (Sep. 2004), S. 5147–5162. DOI: 10.1093/nar/gkh785 (siehe S. 103).

[334] Hongyi Zhou und Yaoqi Zhou. „Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction“. In: *Protein Science* 11.11 (Nov. 2002), S. 2714–2726. DOI: 10.1110/ps.0217002 (siehe S. 103).

[335] Hongyi Zhou und Yaoqi Zhou. „CORRECTION“. In: *Protein Sci.* 12.9 (Sep. 2003), S. 2121–2121. DOI: 10.1002/pro.122121 (siehe S. 103).

[336] Matthew Clark, Richard D. Cramer und Nicole Van Opdenbosch. „Validation of the general purpose tripos 5.2 force field“. In: *J. Comput. Chem.* 10.8 (Dez. 1989), S. 982–1012. DOI: 10.1002/jcc.540100804 (siehe S. 103, 108).

[337] Chi Zhang, Song Liu, Qianqian Zhu und Yaoqi Zhou. „A Knowledge-Based Energy Function for Protein-Ligand, Protein-Protein, and Protein-DNA Complexes“. In: *J. Med. Chem.* 48.7 (Apr. 2005), S. 2325–2335. DOI: 10.1021/jm049314d (siehe S. 104, 117).

[338] James Berger. „The Case for Objective Bayesian Analysis“. In: *Bayesian Analysis* 1 (2006), S. 385–402. DOI: 10.1214/06-ba115 (siehe S. 104).

[339] Timothy A. Robertson und Gabriele Varani. „An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure“. In: *Proteins* 66.2 (Okt. 2006), S. 359–374. DOI: 10.1002/prot.21162 (siehe S. 104, 114).

[340] Yangyu Huang, Shiyong Liu, Dachuan Guo, Lin Li und Yi Xiao. „A novel protocol for three-dimensional structure prediction of RNA-protein complexes“. In: *Sci. Rep.* 3 (Mai 2013). DOI: 10.1038/srep01887 (siehe S. 104, 114).

[341] Z. Yan, L. Guo, L. Hu und J. Wang. „Specificity and affinity quantification of protein-protein interactions“. In: *Bioinformatics* 29.9 (März 2013), S. 1127–1133. DOI: 10.1093/bioinformatics/btt121 (siehe S. 104, 106).

[342] Zhiqiang Yan und Jin Wang. „Optimizing Scoring Function of Protein-Nucleic Acid Interactions with Both Affinity and Specificity“. In: *PLoS ONE* 8.9 (Sep. 2013). Hrsg. von Freddie Salsbury, e74443. DOI: 10.1371/journal.pone.0074443 (siehe S. 104, 106 ff., 110 f., 114, 117).

[343] Sheng-You Huang und Xiaoqin Zou. „An iterative knowledge-based scoring function for protein-protein recognition“. In: *Proteins* 72.2 (Feb. 2008), S. 557–579. DOI: 10.1002/prot.21949 (siehe S. 104, 118).

[344] S.-Y. Huang und X. Zou. „A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method“. In: *Nucleic Acids Research* 42.7 (Jan. 2014), e55–e55. DOI: 10.1093/nar/gku077 (siehe S. 104, 114, 118 ff.).

[345] Dominik Marx und Jürg Hutter. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press, 2009. ISBN: 0521898633 (siehe S. 105).

[346] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly *et al.* „A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules“. In: *J. Am. Chem. Soc.* 117.19 (Mai 1995), S. 5179–5197. DOI: 10.1021/ja00124a002 (siehe S. 105, 121).

[347] Walter R. P. Scott, Philippe H. Hünenberger, Ilario G. Tironi *et al.* „The GROMOS Biomolecular Simulation Program Package“. In: *J. Phys. Chem. A* 103.19 (Mai 1999), S. 3596–3607. DOI: 10.1021/jp984217f (siehe S. 105).

[348] A. D. MacKerell, D. Bashford, M. Bellott *et al.* „All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †“. In: *The Journal of Physical Chemistry B* 102.18 (Apr. 1998), S. 3586–3616. DOI: 10.1021/jp973084f (siehe S. 105).

[349] William L. Jorgensen, David S. Maxwell und Julian Tirado-Rives. „Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids“. In: *J. Am. Chem. Soc.* 118.45 (Jan. 1996), S. 11225–11236. DOI: 10.1021/ja9621760 (siehe S. 105).

[350] Maarten G. Wolf und Gerrit Groenhof. „Evaluating nonpolarizable nucleic acid force fields: A systematic comparison of the nucleobases hydration free energies and chloroform-to-water partition coefficients“. In: *J. Comput. Chem.* 33.28 (Juli 2012), S. 2225–2232. DOI: 10.1002/jcc.23055 (siehe S. 105, 185).

[351] Olgun Guvench und Alexander D. MacKerell. „Comparison of Protein Force Fields for Molecular Dynamics Simulations“. In: *Methods in Molecular Biology*. Springer Science + Business Media, 2008, S. 63–88. DOI: 10.1007/978-1-59745-177-2_4 (siehe S. 105 f.).

[352] Cyril Dominguez, Rolf Boelens und Alexandre M. J. J. Bonvin. „HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information“. In: *J. Am. Chem. Soc.* 125.7 (Feb. 2003), S. 1731–1737. DOI: 10.1021/ja026939x (siehe S. 105 f., 121, 124 f., 189).

[353] Micholas Dean Smith, J. Srinivasa Rao, Elizabeth Segelken und Luis Cruz. „Force-Field Induced Bias in the Structure of Aβ 21–30 : A Comparison of OPLS, AMBER, CHARMM, and GROMOS Force Fields“. In: *Journal of Chemical Information and Modeling* 55.12 (Dez. 2015), S. 2587–2595. DOI: 10.1021/acs.jcim.5b00308 (siehe S. 106).

[354] Thomas E. Cheatham und David A. Case. „Twenty-five years of nucleic acid simulations“. In: *Biopolymers* (Juni 2013). DOI: 10.1002/bip.22331 (siehe S. 106).

[355] Rodrigo Galindo-Murillo, Christina Bergonzo und Thomas E. Cheatham. „Molecular Modeling of Nucleic Acid Structure: Setup and Analysis“. In: *Current Protocols in Nucleic Acid Chemistry*. John Wiley & Sons, Inc., 2001. ISBN: 9780471142706. DOI: 10.1002/0471142700.nc0710s56 (siehe S. 106).

[356] Jin Wang und Gennady M. Verkhivker. „Energy Landscape Theory, Funnels, Specificity, and Optimal Criterion of Biomolecular Binding“. In: *Phys. Rev. Lett.* 90.18 (Mai 2003). DOI: 10.1103/physrevlett.90.188101 (siehe S. 106 f.).

[357] Jin Wang, Xiliang Zheng, Yongliang Yang *et al.* „Quantifying Intrinsic Specificity: A Potential Complement to Affinity in Drug Screening“. In: *Phys. Rev. Lett.* 99.19 (Nov. 2007). DOI: 10.1103/physrevlett.99.198101 (siehe S. 106).

[358] Zhiqiang Yan, Xiliang Zheng, Erkang Wang und Jin Wang. „Thermodynamic and kinetic specificities of ligand binding“. In: *Chemical Science* 4.6 (2013), S. 2387. DOI: 10.1039/c3sc50478f (siehe S. 107).

[359] Zhiqiang Yan und Jin Wang. „Specificity quantification of biomolecular recognition and its implication for drug discovery“. In: *Sci. Rep.* 2 (März 2012). DOI: 10.1038/srep00309 (siehe S. 107).

[360] D. D. Kirsanov, O. N. Zanegina, E. A. Aksianov *et al.* „NPIDB: nucleic acid–protein interaction database“. In: *Nucleic Acids Research* 41.D1 (Nov. 2012), S. D517–D523. DOI: 10.1093/nar/gks1199 (siehe S. 108, 117).

[361] Noel M O'Boyle, Michael Banck, Craig A James *et al.* „Open Babel: An open chemical toolbox“. In: *Journal of Cheminformatics* 3.1 (2011), S. 33. DOI: 10.1186/1758-2946-3-33 (siehe S. 108).

[362] Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis *et al.* „Rosetta3“. In: *Computer Methods, Part C*. Elsevier BV, 2011, S. 545–574. DOI: 10.1016/b978-0-12-381270-4.00019-6 (siehe S. 108, 118).

[363] Sarel J. Fleishman, Andrew Leaver-Fay, Jacob E. Corn *et al.* „RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite“. In: *PLoS ONE* 6.6 (Juni 2011). Hrsg. von Vladimir N. Uversky, e20161. DOI: 10.1371/journal.pone.0020161 (siehe S. 109).

[364] W. A. Koppensteiner und M. J. Sippl. „Knowledge-based potentials–back to the roots.“ In: *Biochemistry (Mosc)* 63.3 (März 1998), S. 247–252. ISSN: 0006-2979 (siehe S. 110).

[365] Zhiqiang Yan. persönliche Kommunikation. 26. Nov. 2015 (siehe S. 112).

[366] Suxin Zheng, Timothy A. Robertson und Gabriele Varani. „A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins“. In: *FEBS Journal* 274.24 (Nov. 2007), S. 6378–6391. DOI: 10.1111/j.1742-4658.2007.06155.x (siehe S. 114).

[367] Hui Lu und Jeffrey Skolnick. „A distance-dependent atomic knowledge-based potential for improved protein structure selection“. In: *Proteins: Structure, Function, and Genetics* 44.3 (2001), S. 223–232. DOI: 10.1002/prot.1087 (siehe S. 114).

[368] Ram Samudrala und John Moult. „An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction“. In: *Journal of Molecular Biology* 275.5 (Feb. 1998), S. 895–916. DOI: 10.1006/jmbi.1997.1479 (siehe S. 114).

[369] Dennis Vitkup, Dagmar Ringe, Martin Karplus und Gregory A. Petsko. „Why protein R-factors are so large: A self-consistent analysis“. In: *Proteins: Structure, Function, and Genetics* 46.4 (Feb. 2002), S. 345–354. DOI: 10.1002/prot.10035 (siehe S. 117).

[370] Z.O. Abu-Faraj. *Handbook of Research on Biomedical Engineering Education and Advanced Bioengineering Learning: Interdisciplinary Concepts: Interdisciplinary Concepts*. Bd. 2. Medical Information Science Reference, 2012. ISBN: 9781466601239 (siehe S. 117).

[371] Xiufeng Yang, Haotian Li, Yangyu Huang und Shiyong Liu. „The dataset for protein-RNA binding affinity“. In: *Protein Science* 22.12 (Nov. 2013), S. 1808–1811. DOI: 10.1002/pro.2383 (siehe S. 117 f.).

[372] Zou Lab. *ITScorePro*. 1. Feb. 2016. URL: http://zoulab.dalton.missouri.edu/resources_itscorepr.html (siehe S. 118, 120).

[373] Rong Chen und Zhiping Weng. „A novel shape complementarity scoring function for protein-protein docking“. In: *Proteins: Structure, Function, and Genetics* 51.3 (Mai 2003), S. 397–408. DOI: 10.1002/prot.10334 (siehe S. 118).

[374] Sheng-You Huang und Xiaoqin Zou. „An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials“. In: *J. Comput. Chem.* 27.15 (2006), S. 1866–1875. DOI: 10.1002/jcc.20504 (siehe S. 118).

[375] M. van Dijk. „Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility“. In: *Nucleic Acids Research* 34.11 (Juni 2006), S. 3317–3325. DOI: 10.1093/nar/gkl412 (siehe S. 120, 124, 189, 193).

[376] D.A. Case, R.M. Betz, W. Botello-Smith *et al. AMBER 2016, University of California, San Francisco*. 2016 (siehe S. 120 f.).

[377] James A. Maier, Carmenza Martinez, Koushik Kasavajhala *et al.* „ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB“. In: *J. Chem. Theory Comput.* 11.8 (Aug. 2015), S. 3696–3713. DOI: 10.1021/acs.jctc.5b00255 (siehe S. 121).

[378] Michael Feig, John Karanicolas und Charles L. Brooks, III. *MMTSB Tool Set, MMTSB NIH Research Resource, The Scripps Research Institute*. 2001 (siehe S. 121).

[379] I. Russo Krauss, A. Pica, V. Napolitano und F. Sica. *X-ray structure of the complex between human alpha thrombin and a duplex/quadruplex 31-mer DNA aptamer*. Jan. 2016. DOI: 10.2210/pdb5cmx/pdb (siehe S. 123).

[380] M. Davlieva, Y. Shamoo und E.P. Nikonowicz. *CRYSTAL STRUCTURE OF BACILLUS ANTHRACIS RIBOSOMAL PROTEIN S8 IN COMPLEX WITH AN RNA APTAMER*. Sep. 2014. DOI: 10.2210/pdb4pdb/pdb (siehe S. 123 f., 126).

[381] K. Ikebukuro. „A novel method of screening thrombin-inhibiting DNA aptamers using an evolution-mimicking algorithm“. In: *Nucleic Acids Research* 33.12 (Juli 2005), e108–e108. DOI: 10.1093/nar/gni108 (siehe S. 123).

[382] A. V. Mazurov, E. V. Titaeva, S. G. Khaspekova *et al.* „Characteristics of a new DNA aptamer, direct inhibitor of thrombin“. In: *Bull. Exp. Biol. Med.* 150.4 (Feb. 2011), S. 422–425 (siehe S. 123).

[383] Irene Russo Krauss, Vera Spiridonova, Andrea Pica, Valeria Napolitano und Filomena Sica. „Different duplex/quadruplex junctions determine the properties of anti-thrombin aptamers with mixed folding“. In: *Nucleic Acids Res* 44.2 (Dez. 2015), S. 983–991. DOI: 10.1093/nar/gkv1384 (siehe S. 123, 126).

[384] D. L. Becker, J. C. Fredenburgh, A. R. Stafford und J. I. Weitz. „Exosites 1 and 2 Are Essential for Protection of Fibrin-bound Thrombin from Heparin-catalyzed Inhibition by Antithrombin and Heparin Cofactor II“. In: *Journal of Biological Chemistry* 274.10 (März 1999), S. 6226–6233. DOI: 10.1074/jbc.274.10.6226 (siehe S. 123).

[385] M. Davlieva, J. Donarski, J. Wang, Y. Shamoo und E. P. Nikonowicz. „Structure analysis of free and bound states of an RNA aptamer against ribosomal protein S8 from Bacillus anthracis“. In: *Nucleic Acids Research* 42.16 (Aug. 2014), S. 10795–10808. DOI: 10.1093/nar/gku743 (siehe S. 124).

[386] E.P. Nikonowicz und J. Wang. *RNA Aptamer for B. anthracis Ribosomal Protein S8*. Dez. 2013. DOI: 10.2210/pdb2lun/pdb (siehe S. 124).

[387] G.C.P. van Zundert, J.P.G.L.M. Rodrigues, M. Trellet *et al.* „The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes“. In: *Journal of Molecular Biology* 428.4 (Feb. 2016), S. 720–725. DOI: 10.1016/j.jmb.2015.09.014 (siehe S. 124, 189).

[388] M. van Dijk und A. M. J. J. Bonvin. „Pushing the limits of what is achievable in protein-DNA docking: benchmarking HADDOCK's performance“. In: *Nucleic Acids Research* 38.17 (Mai 2010), S. 5634–5647. DOI: 10.1093/nar/gkq222 (siehe S. 132).

[389] Sjoerd J. de Vries, Aalt D. J. van Dijk, Mickaël Krzeminski *et al.* „HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets“. In: *Proteins* 69.4 (Sep. 2007), S. 726–733. DOI: 10.1002/prot.21723 (siehe S. 133).

[390] Juan Fernández-Recio, Maxim Totrov und Ruben Abagyan. „Identification of Protein–Protein Interaction Sites from Docking Energy Landscapes“. In: *Journal of Molecular Biology* 335.3 (Jan. 2004), S. 843–865. DOI: 10.1016/j.jmb.2003.10.069 (siehe S. 133).

[391] Schrödinger, LLC. „The PyMOL Molecular Graphics System, Version 1.8“. Nov. 2015 (siehe S. 134, 194).

[392] G. S. Hansman. „Genetic and antigenic diversity among noroviruses“. In: *Journal of General Virology* 87.4 (Apr. 2006), S. 909–919. DOI: 10.1099/vir.0.81532-0 (siehe S. 137).

[393] *Norovirus-Gastroenteritis. RKI-Ratgeber für Ärzte*. Robert Koch-Institut. 26. Juli 2008 (siehe S. 137).

[394] J. L. Adler und R. Zickl. „Winter Vomiting Disease“. In: *Journal of Infectious Diseases* 119.6 (Juni 1969), S. 668–673. DOI: 10.1093/infdis/119.6.668 (siehe S. 137).

[395] Anne M Hutson, Robert L Atmar und Mary K Estes. „Norovirus disease: changing epidemiology and host susceptibility factors“. In: *Trends in Microbiology* 12.6 (Juni 2004), S. 279–287. DOI: 10.1016/j.tim.2004.04.005 (siehe S. 137–140).

[396] A. Z. Kapikian, R. G. Wyatt, R. Dolin *et al.* „Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis“. In: *J. Virol.* 10.5 (Nov. 1972), S. 1075–1081 (siehe S. 137 f., 143).

[397] J. Xi, D. Graham, K. Wang und M. Estes. „Norwalk virus genome cloning and characterization“. In: *Science* 250.4987 (Dez. 1990), S. 1580–1583. DOI: 10.1126/science.2177224 (siehe S. 137).

[398] R. A. Bull, E. T. V. Tu, C. J. McIver, W. D. Rawlinson und P. A. White. „Emergence of a New Norovirus Genotype II.4 Variant Associated with Global Outbreaks of Gastroenteritis“. In: *Journal of Clinical Microbiology* 44.2 (Feb. 2006), S. 327–333. DOI: 10.1128/jcm.44.2.327-333.2006 (siehe S. 138, 141).

[399] Matthew D. Moore, Rebecca M. Goulter und Lee-Ann Jaykus. „Human Norovirus as a Foodborne Pathogen: Challenges and Developments“. In: *Annual Review of Food Science and Technology* 6.1 (Apr. 2015), S. 411–433. DOI: 10.1146/annurev-food-022814-015643 (siehe S. 138 ff., 144).

[400] H. L. Koo, N. Ajami, R. L. Atmar und H. L. DuPont. „Noroviruses: The Principal cause of gastroenteritis worldwide“. In: *Discov Med* 10.50 (Juli 2010), S. 61–70 (siehe S. 138–141, 144).

[401] Jan Vinjé. „Advances in Laboratory Methods for Detection and Typing of Norovirus“. In: *J. Clin. Microbiol.* 53.2 (Juli 2014). Hrsg. von G. V. Doern, S. 373–381. DOI: 10.1128/jcm.01535-14 (siehe S. 139, 143).

[402] E. Duizer. „Laboratory efforts to cultivate noroviruses“. In: *Journal of General Virology* 85.1 (Jan. 2004), S. 79–87. DOI: 10.1099/vir.0.19478-0 (siehe S. 138).

[403] Timothy M. Straub, Kerstin Höner zu Bentrup, Patricia Orosz Coghlan *et al.* „In Vitro Cell Culture Infectivity Assay for Human Noroviruses“. In: *Emerg. Infect. Dis.* 13.3 (März 2007), S. 396–403. DOI: 10.3201/eid1303.060549 (siehe S. 138).

[404] Melissa M. Herbst-Kralovetz, Andrea L. Radtke, Margarita K. Lay *et al.* „Lack of Norovirus Replication and Histo-Blood Group Antigen Expression in 3-Dimensional Intestinal Epithelial Cells“. In: *Emerg. Infect. Dis.* 19.3 (März 2013), S. 431–438. DOI: 10.3201/eid1903.121029 (siehe S. 138).

[405] Efstathia Papafragkou, Joanne Hewitt, Geun Woo Park, Gail Greening und Jan Vinjé. „Challenges of Culturing Human Norovirus in Three-Dimensional Organoid Intestinal Cell Culture Models“. In: *PLoS ONE* 8.6 (Juni 2013). Hrsg. von Amit Kapoor, e63485. DOI: 10.1371/journal.pone.0063485 (siehe S. 138).

[406] Sayaka Takanashi, Linda J. Saif, John H. Hughes *et al.* „Failure of propagation of human norovirus in intestinal epithelial cells with microvilli grown in three-dimensional cultures“. In: *Arch Virol* 159.2 (Aug. 2013), S. 257–266. DOI: 10.1007/s00705-013-1806-4 (siehe S. 138).

[407] M. Souza, M. S. P. Azevedo, K. Jung, S. Cheetham und L. J. Saif. „Pathogenesis and Immune Responses in Gnotobiotic Calves after Infection with the Genogroup II.4-HS66 Strain of Human Norovirus“. In: *Journal of Virology* 82.4 (Nov. 2007), S. 1777–1786. DOI: 10.1128/jvi.01347-07 (siehe S. 138).

[408] B.H.G. Rockx, W.M.J.M. Bogers, J.L. Heeney, G. van Amerongen und M.P.G. Koopmans. „Experimental norovirus infections in non-human primates“. In: *J. Med. Virol.* 75.2 (2004), S. 313–320. DOI: 10.1002/jmv.20273 (siehe S. 138).

[409] K. Bok, G. I. Parra, T. Mitra *et al.* „Chimpanzees as an animal model for human norovirus infection and vaccine development“. In: *Proceedings of the National Academy of Sciences* 108.1 (Dez. 2010), S. 325–330. DOI: 10.1073/pnas.1014577107 (siehe S. 138).

[410] R. Dolin, N. R. Blacklow, H. DuPont *et al.* „Biological properties of Norwalk agent of acute infectious nonbacterial gastroenteritis“. In: *Proc. Soc. Exp. Biol. Med.* 140.2 (Juni 1972), S. 578–583 (siehe S. 138).

[411] R. L. Atmar, A. R. Opekun, M. A. Gilger *et al.* „Determination of the 50% Human Infectious Dose for Norwalk Virus“. In: *Journal of Infectious Diseases* 209.7 (Nov. 2013), S. 1016–1022. DOI: 10.1093/infdis/jit620 (siehe S. 138 ff.).

[412] L. Baert, J. Debevere und M. Uyttendaele. „The efficacy of preservation methods to inactivate foodborne viruses“. In: *International Journal of Food Microbiology* 131.2-3 (Mai 2009), S. 83–94. DOI: 10.1016/j.ijfoodmicro.2009.03.007 (siehe S. 138).

[413] K. Hoelzer, W. Fanaselle, R. Pouillot, J. M. Van Doren und S. Dennis. „Virus Inactivation on Hard Surfaces or in Suspension by Chemical Disinfectants: Systematic Review and Meta-Analysis of Norovirus Surrogates“. In: *j food prot* 76.6 (Juni 2013), S. 1006–1016. DOI: 10.4315/0362-028x.jfp-12-438 (siehe S. 138 f.).

[414] Peter F.M. Teunis, Christine L. Moe, Pengbo Liu *et al.* „Norwalk virus: How infectious is it?“ In: *J. Med. Virol.* 80.8 (2008), S. 1468–1476. DOI: 10.1002/jmv.21237 (siehe S. 139).

[415] Martin C.W. Chan, Joseph J.Y. Sung, Rebecca K.Y. Lam *et al.* „Fecal Viral Load and Norovirus–associated Gastroenteritis“. In: *Emerg. Infect. Dis.* 12.8 (Aug. 2006), S. 1278–1280. DOI: 10.3201/eid1208.060081 (siehe S. 139).

[416] Barry Rockx, Matty de Wit, Harry Vennema *et al.* „Natural History of Human Calicivirus Infection: A Prospective Cohort Study“. In: *Clinical Infectious Diseases* 35.3 (Aug. 2002), S. 246–253. DOI: 10.1086/341408 (siehe S. 139).

[417] Toshio Murata, Noriko Katsushima, Katsumi Mizuta *et al.* „Prolonged Norovirus Shedding in Infants <or=6 Months of Age With Gastroenteritis“. In: *The Pediatric Infectious Disease Journal* 26.1 (Jan. 2007), S. 46–49. DOI: 10.1097/01.inf.0000247102.04997.e0 (siehe S. 139).

[418] David J. Weber, William A. Rutala, Melissa B. Miller, Kirk Huslage und Emily Sickbert-Bennett. „Role of hospital surfaces in the transmission of emerging health care-associated pathogens: Norovirus, Clostridium difficile, and Acinetobacter species“. In: *American Journal of Infection Control* 38.5 (Juni 2010), S25–S33. DOI: 10.1016/j.ajic.2010.04.196 (siehe S. 139).

[419] J. E. MATTHEWS, B. W. DICKEY, R. D. MILLER *et al.* „The epidemiology of published norovirus outbreaks: a review of risk factors associated with attack rate and genogroup“. In: *Epidemiol. Infect.* 140.07 (März 2012), S. 1161–1172. DOI: 10.1017/s0950268812000234 (siehe S. 139).

[420] E. Vega, L. Barclay, N. Gregoricus *et al.* „Genotypic and Epidemiologic Trends of Norovirus Outbreaks in the United States, 2009 to 2013“. In: *Journal of Clinical Microbiology* 52.1 (Okt. 2013), S. 147–155. DOI: 10.1128/jcm.02680-13 (siehe S. 139).

[421] Everardo Vega. „Novel Surveillance Network for Norovirus Gastroenteritis Outbreaks, United States“. In: *Emerg. Infect. Dis.* (Aug. 2011). DOI: 10.3201/eid1708.101837 (siehe S. 139).

[422] Barbara Zanini, Chiara Ricci, Floriana Bandera *et al.* „Incidence of Post-Infectious Irritable Bowel Syndrome and Functional Intestinal Disorders Following a Water-Borne Viral Gastroenteritis Outbreak“. In: *Am J Gastroenterol* 107.6 (Apr. 2012), S. 891–899. DOI: 10.1038/ajg.2012.102 (siehe S. 139).

[423] P. J. Marks, I. B. Vipond, D. Carlisle *et al.* „Evidence for airborne transmission of Norwalk-like virus (NLV) in a hotel restaurant“. In: *Epidemiol. Infect.* 124.3 (Juni 2000), S. 481–487 (siehe S. 139).

[424] Maria Wadl, Kathrin Scherer, Stine Nielsen *et al.* „Food-borne norovirus-outbreak at a military base, Germany, 2009“. In: *BMC Infect Dis* 10.1 (2010), S. 30. DOI: 10.1186/1471-2334-10-30 (siehe S. 139).

[425] P. J. Marks, I. B. Vipond, F. M. Regan *et al.* „A school outbreak of Norwalk-like virus: evidence for airborne transmission“. In: *Epidemiol. Infect.* 131.1 (Aug. 2003), S. 727–736 (siehe S. 139).

[426] Marc-Alain Widdowson, Elaine H. Cramer, Leslie Hadley *et al.* „Outbreaks of Acute Gastroenteritis on Cruise Ships and on Land: Identification of a Predominant Circulating Strain of Norovirus—United States, 2002“. In: *The Journal of Infectious Diseases* 190.1 (Juli 2004), S. 27–36. DOI: 10.1086/420888 (siehe S. 139, 145).

[427] Henry M. Wu, Mary Fornek, Kellogg J. Schwab *et al.* „A Norovirus Outbreak at a Long-Term-Care Facility: The Role of Environmental Surface Contamination“. In: *Infection Control and Hospital Epidemiology* 26.10 (Okt. 2005), S. 802–810. DOI: 10.1086/502497 (siehe S. 139).

[428] E. L. Yee, H. Palacio, R. L. Atmar *et al.* „Widespread Outbreak of Norovirus Gastroenteritis among Evacuees of Hurricane Katrina Residing in a Large “Megashelter” in Houston, Texas: Lessons Learned for Prevention“. In: *Clinical Infectious Diseases* 44.8 (Apr. 2007), S. 1032–1039. DOI: 10.1086/512195 (siehe S. 139).

[429] H. L. Koo, N. J. Ajami, Z. D. Jiang *et al.* „Noroviruses as a Cause of Diarrhea in Travelers to Guatemala, India, and Mexico“. In: *Journal of Clinical Microbiology* 48.5 (März 2010), S. 1673–1676. DOI: 10.1128/jcm.02072-09 (siehe S. 139).

[430] J.-M. Choi, A. M. Hutson, M. K. Estes und B. V. V. Prasad. „Atomic resolution structural characterization of recognition of histo-blood group antigens by Norwalk virus“. In: *Proceedings of the National Academy of Sciences* 105.27 (Juli 2008), S. 9175–9180. DOI: 10.1073/pnas.0803275105 (siehe S. 139).

[431] Robert L. Atmar, Antone R. Opekun, Mark A. Gilger *et al.* „Norwalk Virus Shedding after Experimental Human Infection“. In: *Emerg. Infect. Dis.* 14.10 (Okt. 2008), S. 1553–1557. DOI: 10.3201/eid1410.080117 (siehe S. 140).

[432] H.L. Koo, N.J. Ajami, Z.-D. Jiang, R.L. Atmar und H.L. DuPont. „Norovirus infection as a cause of sporadic healthcare-associated diarrhoea“. In: *Journal of Hospital Infection* 72.2 (Juni 2009), S. 185–187. DOI: 10.1016/j.jhin.2009.03.010 (siehe S. 140).

[433] Stuart S. Kaufman, Nando K. Chatterjee, Meghan E. Fuschino *et al.* „Calicivirus Enteritis in an Intestinal Transplant Recipient“. In: *Am J Transplant* 3.6 (Juni 2003), S. 764–768. DOI: 10.1034/j.1600-6143.2003.00112.x (siehe S. 140).

[434] C. Roddie, J. P. V. Paul, R. Benjamin *et al.* „Allogeneic Hematopoietic Stem Cell Transplantation and Norovirus Gastroenteritis: A Previously Unrecognized Cause of Morbidity“. In: *Clinical Infectious Diseases* 49.7 (Okt. 2009), S. 1061–1068. DOI: 10.1086/605557 (siehe S. 140).

[435] Anne M. Hutson, Robert L. Atmar, David Y. Graham und Mary K. Estes. „Norwalk Virus Infection and Disease Is Associated with ABO Histo–Blood Group Type“. In: *The Journal of Infectious Diseases* 185.9 (Mai 2002), S. 1335–1337. DOI: 10.1086/339883 (siehe S. 140).

[436] Lisa Lindesmith, Christine Moe, Severine Marionneau *et al.* „Human susceptibility and resistance to Norwalk virus infection“. In: *Nature Medicine* 9.5 (Apr. 2003), S. 548–553. DOI: 10.1038/nm860 (siehe S. 140).

[437] J. Swanstrom, L. C. Lindesmith, E. F. Donaldson, B. Yount und R. S. Baric. „Characterization of Blockade Antibody Responses in GII.2.1976 Snow Mountain Virus-Infected Subjects“. In: *Journal of Virology* 88.2 (Okt. 2013), S. 829–837. DOI: 10.1128/jvi.02793-13 (siehe S. 140).

[438] P. C. Johnson, J. J. Mathewson, H. L. DuPont und H. B. Greenberg. „Multiple-Challenge Study of Host Susceptibility to Norwalk Gastroenteritis in US Adults“. In: *Journal of Infectious Diseases* 161.1 (Jan. 1990), S. 18–21. DOI: 10.1093/infdis/161.1.18 (siehe S. 140).

[439] R. A. Bull, J.-S. Eden, F. Luciani *et al.* „Contribution of Intra- and Interhost Dynamics to Norovirus Evolution“. In: *Journal of Virology* 86.6 (Dez. 2011), S. 3219–3229. DOI: 10.1128/jvi.06712-11 (siehe S. 140).

[440] Robert L. Atmar, David I. Bernstein, Clayton D. Harro *et al.* „Norovirus Vaccine against Experimental Human Norwalk Virus Illness“. In: *New England Journal of Medicine* 365.23 (Dez. 2011), S. 2178–2187. DOI: 10.1056/nejmoa1101245 (siehe S. 140).

[441] Xiuren Zhang, Norene A. Buehner, Anne M. Hutson, Mary K. Estes und Hugh S. Mason. „Tomato is a highly effective vehicle for expression and oral immunization with Norwalk virus capsid protein“. In: *Plant Biotechnology Journal* 4.4 (Juli 2006), S. 419–432. DOI: 10.1111/j.1467-7652.2006.00191.x (siehe S. 140).

[442] Carol O. Tacket, Hugh S. Mason, Genevieve Losonsky *et al.* „Human Immune Responses to a Novel Norwalk Virus Vaccine Delivered in Transgenic Potatoes“. In: *The Journal of Infectious Diseases* 182.1 (Juli 2000), S. 302–305. DOI: 10.1086/315653 (siehe S. 140).

[443] Danish M. Siddiq, Hoonmo L. Koo, Javier A. Adachi und George M. Viola. „Norovirus gastroenteritis successfully treated with nitazoxanide“. In: *Journal of Infection* 63.5 (Nov. 2011), S. 394–397. DOI: 10.1016/j.jinf.2011.08.002 (siehe S. 140).

[444] Zain Chagla, Jaclyn Quirt, Kevin Woodward, John Neary und Candace Rutherford. „Chronic norovirus infection in a transplant patient successfully treated with enterally administered immune globulin“. In: *Journal of Clinical Virology* 58.1 (Sep. 2013), S. 306–308. DOI: 10.1016/j.jcv.2013.06.009 (siehe S. 140).

[445] Stuart S. Kaufman, Kim Y. Green und Brent E. Korba. „Treatment of norovirus infections: Moving antivirals from the bench to the bedside“. In: *Antiviral Research* 105 (Mai 2014), S. 80–91. DOI: 10.1016/j.antiviral.2014.02.012 (siehe S. 140).

[446] Armando Arias, Edward Emmott, Surender Vashist und Ian Goodfellow. „Progress towards the prevention and treatment of norovirus infections“. In: *Future Microbiology* 8.11 (Nov. 2013), S. 1475–1487. DOI: 10.2217/fmb.13.109 (siehe S. 140).

[447] J. D. GREIG und M. B. LEE. „A review of nosocomial norovirus outbreaks: infection control interventions found effective“. In: *Epidemiol. Infect.* 140.07 (Jan. 2012), S. 1151–1160. DOI: 10.1017/s0950268811002731 (siehe S. 140).

[448] *Preventing Norovirus Infection.* 10. Dez. 2015 (siehe S. 140).

[449] J.P. Harris, B.A. Lopman und S.J. O'Brien. „Infection control measures for norovirus: a systematic review of outbreaks in semi-enclosed settings". In: *Journal of Hospital Infection* 74.1 (Jan. 2010), S. 1–9. DOI: 10.1016/j.jhin.2009.07.025 (siehe S. 140).

[450] C. S. Manuel, M. D. Moore und L. A. Jaykus. „Destruction of the Capsid and Genome of GII.4 Human Norovirus Occurs during Exposure to Metal Alloys Containing Copper". In: *Applied and Environmental Microbiology* 81.15 (Mai 2015). Hrsg. von K. E. Wommack, S. 4940–4946. DOI: 10.1128/aem.00388-15 (siehe S. 141).

[451] Anna D. Koromyslova, Peter A. White und Grant S. Hansman. „Treatment of norovirus particles with citrate". In: *Virology* 485 (Nov. 2015), S. 199–204. DOI: 10.1016/j.virol.2015.07.009 (siehe S. 141).

[452] B. V. V. Prasad. „X-ray Crystallographic Structure of the Norwalk Virus Capsid". In: *Science* 286.5438 (Okt. 1999), S. 287–290. DOI: 10.1126/science.286.5438.287 (siehe S. 141 f.).

[453] B. V. Venkataram Prasad, M. E. Hardy und M. K. Estes. „Structural Studies of Recombinant Norwalk Capsids". In: *The Journal of Infectious Diseases* 181.s2 (Mai 2000), S317–S321. DOI: 10.1086/315576 (siehe S. 141).

[454] R. Chen, J. D. Neill, M. K. Estes und B. V. V. Prasad. „X-ray structure of a native calicivirus: Structural insights into antigenic diversity and host specificity". In: *Proceedings of the National Academy of Sciences* 103.21 (Mai 2006), S. 8048–8053. DOI: 10.1073/pnas.0600421103 (siehe S. 141).

[455] Rong Chen, John D. Neill und B.V. Venkataram Prasad. „Crystallization and preliminary crystallographic analysis of San Miguel sea lion virus: An animal calicivirus". In: *Journal of Structural Biology* 141.2 (Feb. 2003), S. 143–148. DOI: 10.1016/s1047-8477(02)00583-x (siehe S. 141).

[456] R. J. Ossiboff, Y. Zhou, P. J. Lightfoot, B. V. V. Prasad und J. S. L. Parker. „Conformational Changes in the Capsid of a Calicivirus upon Interaction with Its Functional Receptor". In: *Journal of Virology* 84.11 (März 2010), S. 5550–5564. DOI: 10.1128/jvi.02371-09 (siehe S. 141).

[457] S. Vongpunsawad, B. V. Venkataram Prasad und M. K. Estes. „Norwalk Virus Minor Capsid Protein VP2 Associates within the VP1 Shell Domain". In: *Journal of Virology* 87.9 (Feb. 2013), S. 4818–4825. DOI: 10.1128/jvi.03508-12 (siehe S. 141).

[458] B.V. Prasad, M.E. Hardy, T. Dokland *et al. CRYSTAL STRUCTURE ANALYSIS OF NORWALK VIRUS CAPSID*. Mai 2001. DOI: 10.2210/pdb1ihm/pdb (siehe S. 142, 181 f., 199).

[459] B. V. Venkataram Prasad und Michael F. Schmid. „Principles of Virus Structural Organization". In: *Viral Molecular Machines*. Springer Science + Business Media, Nov. 2011, S. 17–47. DOI: 10.1007/978-1-4614-0980-9_3 (siehe S. 142).

[460] H. Shirato, S. Ogawa, H. Ito *et al.* „Noroviruses Distinguish between Type 1 and Type 2 Histo-Blood Group Antigens for Binding". In: *Journal of Virology* 82.21 (Aug. 2008), S. 10756–10767. DOI: 10.1128/jvi.00802-08 (siehe S. 142).

[461] Kosuke Murakami, Chie Kurihara, Tomoichiro Oka *et al.* „Norovirus Binding to Intestinal Epithelial Cells Is Independent of Histo-Blood Group Antigens". In: *PLoS ONE* 8.6 (Juni 2013). Hrsg. von Karol Sestak, e66534. DOI: 10.1371/journal.pone.0066534 (siehe S. 142).

[462] S. F. Ausar. „Conformational Stability and Disassembly of Norwalk Virus-like Particles: EFFECT OF pH AND TEMPERATURE“. In: *Journal of Biological Chemistry* 281.28 (Mai 2006), S. 19478–19488. DOI: 10.1074/jbc.m603313200 (siehe S. 142).

[463] JONATHAN E. KAPLAN. „Epidemiology of Norwalk Gastroenteritis and the Role of Norwalk Virus in Outbreaks of Acute Nonbacterial Gastroenteritis“. In: *Annals of Internal Medicine* 96.6_part_1 (Juni 1982), S. 756. DOI: 10.7326/0003-4819-96-6-756 (siehe S. 143).

[464] A.F Richards, B Lopman, A Gunn *et al.* „Evaluation of a commercial ELISA for detecting Norwalk-like virus antigen in faeces“. In: *Journal of Clinical Virology* 26.1 (Jan. 2003), S. 109–115. DOI: 10.1016/s1386-6532(02)00267-6 (siehe S. 143).

[465] Katia Ambert-Balay und Pierre Pothier. „Evaluation of 4 immunochromatographic tests for rapid detection of norovirus in faecal samples“. In: *Journal of Clinical Virology* 56.3 (März 2013), S. 278–282. DOI: 10.1016/j.jcv.2012.11.001 (siehe S. 143).

[466] V. Costantini, L. Grenz, A. Fritzinger *et al.* „Diagnostic Accuracy and Analytical Sensitivity of IDEIA Norovirus Assay for Routine Screening of Human Norovirus“. In: *Journal of Clinical Microbiology* 48.8 (Juni 2010), S. 2770–2778. DOI: 10.1128/jcm.00654-10 (siehe S. 143).

[467] J. J. Gray, E. Kohli, F. M. Ruggeri *et al.* „European Multicenter Evaluation of Commercial Enzyme Immunoassays for Detecting Norovirus Antigen in Fecal Samples“. In: *Clinical and Vaccine Immunology* 14.10 (Aug. 2007), S. 1349–1355. DOI: 10.1128/cvi.00214-07 (siehe S. 143).

[468] H Vennema, E de Bruin und M Koopmans. „Rational optimization of generic primers used for Norwalk-like virus detection by reverse transcriptase polymerase chain reaction“. In: *Journal of Clinical Virology* 25.2 (Aug. 2002), S. 233–235. DOI: 10.1016/s1386-6532(02)00126-9 (siehe S. 143).

[469] Kazuhiko Katayama, Haruko Shirato-Horikoshi, Shigeyuki Kojima *et al.* „Phylogenetic Analysis of the Complete Genome of 18 Norwalk-like Viruses“. In: *Virology* 299.2 (Aug. 2002), S. 225–239. DOI: 10.1006/viro.2002.1568 (siehe S. 143).

[470] T. Miura, S. Parnaudeau, M. Grodzki *et al.* „Environmental Detection of Genogroup I, II, and IV Noroviruses by Using a Generic Real-Time Reverse Transcription-PCR Assay“. In: *Applied and Environmental Microbiology* 79.21 (Aug. 2013), S. 6585–6592. DOI: 10.1128/aem.02112-13 (siehe S. 143).

[471] Hengjia Ni, Suxia Zhang, Xinghua Ding *et al.* „Determination of Enrofloxacin in Bovine Milk by a Novel Single-Stranded DNA Aptamer Chemiluminescent Enzyme Immunoassay“. In: *Analytical Letters* 47.17 (Sep. 2014), S. 2844–2856. DOI: 10.1080/00032719.2014.924009 (siehe S. 144).

[472] Vincent J. B. Ruigrok, Mark Levisson, Michel H. M. Eppink, Hauke Smidt und John van der Oost. „Alternative affinity tools: more attractive than antibodies?“ In: *Biochem. J.* 436.1 (Mai 2011), S. 1–13. DOI: 10.1042/bj20101860 (siehe S. 144).

[473] John-Sebastian Eden, Joanne Hewitt, Kun Lee Lim *et al.* „The emergence and evolution of the novel epidemic norovirus GII.4 variant Sydney 2012“. In: *Virology* 450-451 (Feb. 2014), S. 106–113. DOI: 10.1016/j.virol.2013.12.005 (siehe S. 144).

[474] Makoto Kumazaki und Shuzo Usuku. „Genetic Analysis of Norovirus GII.4 Variant Strains Detected in Outbreaks of Gastroenteritis in Yokohama, Japan, from the 2006-2007 to the 2013-2014 Seasons“. In: *PLOS ONE* 10.11 (Nov. 2015). Hrsg. von Patrick CY Woo, e0142568. DOI: 10.1371/journal.pone.0142568 (siehe S. 144 f.).

[475] Eric F. Donaldson, Lisa C. Lindesmith, Anna D. LoBue und Ralph S. Baric. „Viral shape-shifting: norovirus evasion of the human immune system“. In: *Nature Reviews Microbiology* 8.3 (Feb. 2010), S. 231–241. DOI: 10.1038/nrmicro2296 (siehe S. 144).

[476] J. L. Cannon, L. C. Lindesmith, E. F. Donaldson *et al.* „Herd Immunity to GII.4 Noroviruses Is Supported by Outbreak Patient Sera“. In: *Journal of Virology* 83.11 (März 2009), S. 5363–5374. DOI: 10.1128/jvi.02518-08 (siehe S. 144).

[477] M. F Boni, J. R Gog, V. Andreasen und M. W Feldman. „Epidemic dynamics and antigenic evolution in a single season of influenza A“. In: *Proceedings of the Royal Society B: Biological Sciences* 273.1592 (Juni 2006), S. 1307–1316. DOI: 10.1098/rspb.2006.3466 (siehe S. 144).

[478] F. Sanger, S. Nicklen und A. R. Coulson. „DNA sequencing with chain-terminating inhibitors“. In: *Proc. Natl. Acad. Sci. U.S.A.* 74.12 (Dez. 1977), S. 5463–5467 (siehe S. 145).

[479] Michael L. Metzker. „Sequencing technologies — the next generation“. In: *Nat Rev Genet* 11.1 (Dez. 2009), S. 31–46. DOI: 10.1038/nrg2626 (siehe S. 145).

[480] Meltem Avci-Adali, Angel Paul, Nadj Wilhelm, Gerhard Ziemer und Hans Peter Wendel. „Upgrading SELEX Technology by Using Lambda Exonuclease Digestion for Single-Stranded DNA Generation“. In: *Molecules* 15.1 (Dez. 2009), S. 1–11. DOI: 10.3390/molecules15010001 (siehe S. 146).

[481] Chengwei Luo, Despina Tsementzi, Nikos Kyrpides, Timothy Read und Konstantinos T. Konstantinidis. „Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample“. In: *PLoS ONE* 7.2 (Feb. 2012). Hrsg. von Francisco Rodriguez-Valera, e30087. DOI: 10.1371/journal.pone.0030087 (siehe S. 149).

[482] C. J. Keylock. „Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy“. In: *Oikos* 109.1 (Apr. 2005), S. 203–207. DOI: 10.1111/j.0030-1299.2005.13735.x (siehe S. 150).

[483] Priyabrata Pattnaik. „Surface Plasmon Resonance: Applications in Understanding Receptor–Ligand Interaction“. In: *Applied Biochemistry and Biotechnology* 126.2 (2005), S. 079–092. DOI: 10.1385/abab:126:2:079 (siehe S. 152).

[484] Hana Šípová und Jiří Homola. „Surface plasmon resonance sensing of nucleic acids: A review“. In: *Analytica Chimica Acta* 773 (Apr. 2013), S. 9–23. DOI: 10.1016/j.aca.2012.12.040 (siehe S. 152).

[485] Martin Specht, Johannes Pedarnig, Wolfgang Heckl und Theodor Hänsch. „Das Plasmonenmikroskop“. In: *Physik in unserer Zeit* 24.4 (1993), S. 176–179. DOI: 10.1002/piuz.19930240409 (siehe S. 152).

[486] Anja Henseleit, Stefan Schmieder, Thomas Bley *et al.* „Oberflächenfunktionalisierung von Goldschichten zur gerichteten Immobilisierung von Biomolekülen“. In: (2012) (siehe S. 152).

[487] Rebecca L. Rich und David G. Myszka. „Survey of the year 2007 commercial optical biosensor literature“. In: *J. Mol. Recognit.* 21.6 (Nov. 2008), S. 355–400. DOI: 10.1002/jmr.928 (siehe S. 153).

[488] C. Schudoma, P. May, V. Nikiforova und D. Walther. „Sequence-structure relationships in RNA loops: establishing the basis for loop homology modeling“. In: *Nucleic Acids Research* 38.3 (Nov. 2009), S. 970–980. DOI: 10.1093/nar/gkp1010 (siehe S. 158, 162).

[489] J. Hoinka, E. Zotenko, A. Friedman, Z. E. Sauna und T. M. Przytycka. „Identification of sequence-structure RNA binding motifs for SELEX-derived aptamers“. In: *Bioinformatics* 28.12 (Juni 2012), S. i215–i223. DOI: 10.1093/bioinformatics/bts210 (siehe S. 158, 162).

[490] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen *et al.* „ViennaRNA Package 2.0“. In: *Algorithms Mol Biol* 6.1 (2011), S. 26. DOI: 10.1186/1748-7188-6-26 (siehe S. 158).

[491] D. H. Turner und D. H. Mathews. „NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure“. In: *Nucleic Acids Research* 38.Database (Okt. 2009), S. D280–D282. DOI: 10.1093/nar/gkp892 (siehe S. 158).

[492] John SantaLucia und Donald Hicks. „The Thermodynamics of DNA Structural Motifs“. In: *Annual Review of Biophysics and Biomolecular Structure* 33.1 (Juni 2004), S. 415–440. DOI: 10.1146/annurev.biophys.32.110601.141800 (siehe S. 158).

[493] D. H. Mathews, M. D. Disney, J. L. Childs *et al.* „Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure“. In: *Proceedings of the National Academy of Sciences* 101.19 (Mai 2004), S. 7287–7292. DOI: 10.1073/pnas.0401799101 (siehe S. 158).

[494] Wayne Dawson, Toshikuni Takai, Nobuharu Ito, Kentaro Shimizu und Gota Kawai. „A new entropy model for RNA: part III. Is the folding free energy landscape of RNA funnel shaped?“ In: *J Nucleic Acids Invest* 5.1 (Nov. 2014). DOI: 10.4081/jnai.2014.2652 (siehe S. 158).

[495] Alexander Churkin, Lina Weinbrand und Danny Barash. „Free Energy Minimization to Predict RNA Secondary Structures and Computational RNA Design“. In: *Methods in Molecular Biology*. Springer Science + Business Media, Dez. 2014, S. 3–16. DOI: 10.1007/978-1-4939-2291-8_1 (siehe S. 158).

[496] Ivo L. Hofacker. „Energy-Directed RNA Structure Prediction“. In: *Methods in Molecular Biology*. Springer Science + Business Media, Dez. 2013, S. 71–84. DOI: 10.1007/978-1-62703-709-9_4 (siehe S. 158).

[497] D. H. Mathews, M. E. Burkard, S. M. Freier, J. R. Wyatt und D. H. Turner. „Predicting oligonucleotide affinity to nucleic acid targets“. In: *RNA* 5.11 (Nov. 1999), S. 1458–1469 (siehe S. 158).

[498] I MIKLOS, I MEYER und B NAGY. „Moments of the Boltzmann distribution for RNA secondary structures“. In: *Bulletin of Mathematical Biology* 67.5 (Sep. 2005), S. 1031–1047. DOI: 10.1016/j.bulm.2004.12.003 (siehe S. 158).

[499] Gregory J. Connell, Mali Illangesekare und Michael Yarus. „Three small ribooligonucleotides with specific arginine sites“. In: *Biochemistry* 32.21 (Juni 1993), S. 5497–5502. DOI: 10.1021/bi00072a002 (siehe S. 159).

[500] M. LEGIEWICZ. „Size, constant sequences, and optimal selection“. In: *RNA* 11.11 (Nov. 2005), S. 1701–1709. DOI: 10.1261/rna.2161305 (siehe S. 159).

[501] C. LOZUPONE. „Selection of the simplest RNA that binds isoleucine“. In: *RNA* 9.11 (Nov. 2003), S. 1315–1322. DOI: 10.1261/rna.5114503 (siehe S. 159).

[502] I. Majerfeld und M. Yarus. „Isoleucine:RNA sites with associated coding sequences“. In: *RNA* 4.4 (Apr. 1998), S. 471–478 (siehe S. 159).

[503] F. Jarosch. „In vitro selection using a dual RNA library that allows primerless selection“. In: *Nucleic Acids Research* 34.12 (Juli 2006), e86–e86. DOI: 10.1093/nar/gkl463 (siehe S. 159).

[504] Matthew C. Cowperthwaite und Andrew D. Ellington. „Bioinformatic Analysis of the Contribution of Primer Sequences to Aptamer Structures“. In: *J Mol Evol* 67.1 (Juli 2008), S. 95–102. DOI: 10.1007/s00239-008-9130-4 (siehe S. 159).

[505] K. A. Dill und J. L. MacCallum. „The Protein-Folding Problem, 50 Years On“. In: *Science* 338.6110 (Nov. 2012), S. 1042–1046. DOI: 10.1126/science.1219021 (siehe S. 179).

[506] Justin L. MacCallum, Alberto Pérez, Michael J. Schnieders *et al.* „Assessment of protein structure refinement in CASP9“. In: *Proteins* 79.S10 (2011), S. 74–90. DOI: 10.1002/prot.23131 (siehe S. 179).

[507] Justin L. MacCallum, Lan Hua, Michael J. Schnieders *et al.* „Assessment of the protein-structure refinement category in CASP8“. In: *Proteins* 77.S9 (2009), S. 66–80. DOI: 10.1002/prot.22538 (siehe S. 179).

[508] Conor R. Caffrey, Lenka Placha, Cyril Barinka *et al.* „Homology modeling and SAR analysis of Schistosoma japonicum cathepsin D (SjCD) with statin inhibitors identify a unique active site steric barrier with potential for the design of specific inhibitors“. In: *Biological Chemistry* 386.4 (Jan. 2005). DOI: 10.1515/bc.2005.041 (siehe S. 179).

[509] Gordon A. Wells, Lyn-Marie Birkholtz, Fourie Joubert, Rolf D. Walter und Abraham I. Louw. „Novel properties of malarial S-adenosylmethionine decarboxylase as revealed by structural modelling“. In: *Journal of Molecular Graphics and Modelling* 24.4 (Jan. 2006), S. 307–318. DOI: 10.1016/j.jmgm.2005.09.011 (siehe S. 179).

[510] J. Peng und J. Xu. „Low-homology protein threading“. In: *Bioinformatics* 26.12 (Juni 2010), S. i294–i300. DOI: 10.1093/bioinformatics/btq192 (siehe S. 179).

[511] D. T. Jones, W. R. Taylort und J. M. Thornton. „A new approach to protein fold recognition“. In: *Nature* 358.6381 (Juli 1992), S. 86–89. DOI: 10.1038/358086a0 (siehe S. 179).

[512] J. Bowie, R Luthy und D Eisenberg. „A method to identify protein sequences that fold into a known three-dimensional structure“. In: *Science* 253.5016 (Juli 1991), S. 164–170. DOI: 10.1126/science.1853201 (siehe S. 179).

[513] George A. Khoury, James Smadbeck, Chris A. Kieslich und Christodoulos A. Floudas. „Protein folding and de novo protein design for biotechnological applications“. In: *Trends in Biotechnology* 32.2 (Feb. 2014), S. 99–109. DOI: 10.1016/j.tibtech.2013.10.008 (siehe S. 179 f.).

[514] Yang Zhang. „Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10“. In: *Proteins* 82 (Aug. 2013), S. 175–187. DOI: 10.1002/prot.24341 (siehe S. 179, 182).

[515] Keehyoung Joo, Juyong Lee, Sangjin Sim *et al.* „Protein structure modeling for CASP10 by multiple layers of global optimization“. In: *Proteins* 82 (Okt. 2013), S. 188–195. DOI: 10.1002/prot.24397 (siehe S. 179).

[516] D. T. Jones, D. W. A. Buchan, D. Cozzetto und M. Pontil. „PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments“. In: *Bioinformatics* 28.2 (Nov. 2011), S. 184–190. DOI: 10.1093/bioinformatics/btr638 (siehe S. 179).

[517] R. Rajgaria, S. R. McAllister und C. A. Floudas. „Towards accurate residue-residue hydrophobic contact prediction for α helical proteins via integer linear optimization“. In: *Proteins* 74.4 (März 2009), S. 929–947. DOI: 10.1002/prot.22202 (siehe S. 179).

[518] R. Rajgaria, Y. Wei und C. A. Floudas. „Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD“. In: *Proteins* (2010), NA–NA. DOI: 10.1002/prot.22696 (siehe S. 179).

[519] David E. Kim, Frank DiMaio, Ray Yu-Ruei Wang, Yifan Song und David Baker. „One contact for every twelve residues allows robust and accurate topology-level protein structure modeling“. In: *Proteins* 82 (Sep. 2013), S. 208–218. DOI: 10.1002/prot.24374 (siehe S. 180).

[520] K. Lindorff-Larsen, S. Piana, R. O. Dror und D. E. Shaw. „How Fast-Folding Proteins Fold“. In: *Science* 334.6055 (Okt. 2011), S. 517–520. DOI: 10.1126/science.1208351 (siehe S. 180).

[521] David E. Shaw, Kevin J. Bowers, Edmond Chow *et al.* „Millisecond-scale molecular dynamics simulations on Anton“. In: *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09*. Association for Computing Machinery (ACM), 2009. DOI: 10.1145/1654059.1654126 (siehe S. 180).

[522] Valerio Mariani, Florian Kiefer, Tobias Schmidt, Juergen Haas und Torsten Schwede. „Assessment of template based protein structure predictions in CASP9“. In: *Proteins* 79.S10 (2011), S. 37–58. DOI: 10.1002/prot.23177 (siehe S. 180).

[523] Yuanpeng J. Huang, Binchen Mao, James M. Aramini und Gaetano T. Montelione. „Assessment of template-based protein structure predictions in CASP10“. In: *Proteins* 82 (Jan. 2014), S. 43–56. DOI: 10.1002/prot.24488 (siehe S. 180).

[524] Jianyi Yang, Wenxuan Zhang, Baoji He *et al.* „Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade“. In: *Proteins* (Sep. 2015), n/a–n/a. DOI: 10.1002/prot.24918 (siehe S. 180).

[525] Ambrish Roy, Alper Kucukural und Yang Zhang. „I-TASSER: a unified platform for automated protein structure and function prediction“. In: *Nat Protoc* 5.4 (März 2010), S. 725–738. DOI: 10.1038/nprot.2010.5 (siehe S. 180, 182).

[526] S. Wu und Y. Zhang. „LOMETS: A local meta-threading-server for protein structure prediction“. In: *Nucleic Acids Research* 35.10 (Apr. 2007), S. 3375–3382. DOI: 10.1093/nar/gkm251 (siehe S. 181).

[527] D. Luque, J.M. Gonzalez, J. Gomez-Blanco *et al. Rabbit Hemorrhagic Disease Virus (RHDV) capsid protein*. Mai 2012. DOI: 10.2210/pdb3zue/pdb (siehe S. 181).

[528] Z. Xue, D. Xu, Y. Wang und Y. Zhang. „ThreaDom: extracting protein domain boundary information from multiple threading alignments“. In: *Bioinformatics* 29.13 (Juni 2013), S. i247–i256. DOI: 10.1093/bioinformatics/btt209 (siehe S. 181).

[529] The UniProt Consortium. „UniProt: a hub for protein information“. In: *Nucleic Acids Research* 43.D1 (Okt. 2014), S. D204–D212. DOI: 10.1093/nar/gku989 (siehe S. 182).

[530] S. Shanker, J.-M. Choi, B. Sankaran *et al. Structural characterization of a GII.4 2004 norovirus variant (TCH05) bound to A trisaccharide*. Juli 2011. DOI: 10.2210/pdb3sld/pdb (siehe S. 182).

[531] Jianyi Yang, Renxiang Yan, Ambrish Roy *et al.* „The I-TASSER Suite: protein structure and function prediction“. In: *Nature Methods* 12.1 (Dez. 2014), S. 7–8. DOI: 10.1038/nmeth.3213 (siehe S. 182).

[532] Jianyi Yang und Yang Zhang. „I-TASSER server: new development for protein structure and function predictions“. In: *Nucleic Acids Res* 43.W1 (Apr. 2015), W174–W181. DOI: 10.1093/nar/gkv342 (siehe S. 182).

[533] Yang Zhang. „I-TASSER server for protein 3D structure prediction“. In: *BMC Bioinformatics* 9.1 (2008), S. 40. DOI: 10.1186/1471-2105-9-40 (siehe S. 182).

[534] Yang Zhang und Jeffrey Skolnick. „Scoring function for automated assessment of protein structure template quality“. In: *Proteins* 57.4 (2004), S. 702–710. DOI: 10.1002/prot.20264 (siehe S. 182).

[535] Jianyi Yang, Yan Wang und Yang Zhang. „ResQ: An Approach to Unified Estimation of B-Factor and Residue-Specific Error in Protein Structure Prediction“. In: *Journal of Molecular Biology* 428.4 (Feb. 2016), S. 693–701. DOI: 10.1016/j.jmb.2015.09.024 (siehe S. 182 f.).

[536] I. N. Shindyalov und P. E. Bourne. „Protein structure alignment by incremental combinatorial extension (CE) of the optimal path“. In: *Protein Engineering Design and Selection* 11.9 (Sep. 1998), S. 739–747. DOI: 10.1093/protein/11.9.739 (siehe S. 183).

[537] Jian Zhang, Yu Liang und Yang Zhang. „Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling“. In: *Structure* 19.12 (Dez. 2011), S. 1784–1795. DOI: 10.1016/j.str.2011.09.022 (siehe S. 183).

[538] Christian Laing und Tamar Schlick. „Computational approaches to RNA structure prediction, analysis, and design“. In: *Current Opinion in Structural Biology* 21.3 (Juni 2011), S. 306–318. DOI: 10.1016/j.sbi.2011.03.015 (siehe S. 185).

[539] Yunjie Zhao, Yangyu Huang, Zhou Gong *et al.* „Automated and fast building of three-dimensional RNA structures“. In: *Sci. Rep.* 2 (Okt. 2012). DOI: 10.1038/srep00734 (siehe S. 185).

[540] Zhichao Miao, Ryszard W. Adamiak, Marc-Frédérick Blanchet *et al.* „RNA-PuzzlesRound II: assessment of RNA structure prediction programs applied to three large RNA structures“. In: *RNA* 21.6 (Apr. 2015), S. 1066–1084. DOI: 10.1261/rna.049502.114 (siehe S. 185).

[541] M. Rother, K. Rother, T. Puton und J. M. Bujnicki. „ModeRNA: a tool for comparative modeling of RNA 3D structure“. In: *Nucleic Acids Research* 39.10 (Feb. 2011), S. 4007–4022. DOI: 10.1093/nar/gkq1320 (siehe S. 185).

[542] S. C. Flores, Y. Wan, R. Russell und R. B. Altman. „Predicting RNA structure by multiple template homology modeling“. In: *Pac Symp Biocomput* (2010), S. 216–227 (siehe S. 185).

[543] Lili Dong, Qiwen Tan, Wei Ye *et al.* „Screening and Identifying a Novel ssDNA Aptamer against Alpha-fetoprotein Using CE-SELEX“. In: *Sci. Rep.* 5 (Okt. 2015), S. 15552. DOI: 10.1038/srep15552 (siehe S. 185).

[544] Iman Jeddi und Leonor Saiz. „Structure Prediction and 3D Modeling of Single Stranded DNA from Sequence for Aptamer-Based Biosensors“. In: *Biophysical Journal* 110.3 (Feb. 2016), 333a. DOI: 10.1016/j.bpj.2015.11.1792 (siehe S. 185).

[545] Rhiju Das, John Karanicolas und David Baker. „Atomic accuracy in predicting and designing noncanonical RNA structure“. In: *Nature Methods* 7.4 (Feb. 2010), S. 291–294. DOI: 10.1038/nmeth.1433 (siehe S. 185).

[546] S. Sharma, F. Ding und N. V. Dokholyan. „iFoldRNA: three-dimensional RNA structure prediction and folding“. In: *Bioinformatics* 24.17 (Juni 2008), S. 1951–1952. DOI: 10.1093/bioinformatics/btn328 (siehe S. 185).

[547] Andrey Krokhotin, Kevin Houlihan und Nikolay V. Dokholyan. „iFoldRNA v2: folding RNA with constraints.“ In: *Bioinformatics* 31.17 (Apr. 2015), S. 2891–2893. DOI: 10.1093/bioinformatics/btv221 (siehe S. 185).

[548] Michal J. Boniecki, Grzegorz Lach, Wayne K. Dawson *et al.* „SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction“. In: *Nucleic Acids Res* 44.7 (Dez. 2015), e63–e63. DOI: 10.1093/nar/gkv1479 (siehe S. 185).

[549] M. van Dijk und A. M. J. J. Bonvin. „3D-DART: a DNA structure modelling server“. In: *Nucleic Acids Research* 37.Web Server (Mai 2009), W235–W239. DOI: 10.1093/nar/gkp287 (siehe S. 185).

[550] Alex Tek, Andrei A. Korostelev und Samuel Coulbourn Flores. „MMB-GUI: a fast morphing method demonstrates a possible ribosomal tRNA translocation trajectory“. In: *Nucleic Acids Res* 44.1 (Dez. 2015), S. 95–105. DOI: 10.1093/nar/gkv1457 (siehe S. 185).

[551] Xavier Periole und Siewert-Jan Marrink. „The Martini Coarse-Grained Force Field“. In: *Methods in Molecular Biology*. Springer Science + Business Media, Aug. 2012, S. 533–565. DOI: 10.1007/978-1-62703-017-5_20 (siehe S. 185 f., 188).

[552] Jaakko J. Uusitalo, Helgi I. Ingólfsson, Parisa Akhshi, D. Peter Tieleman und Siewert J. Marrink. „Martini Coarse-Grained Force Field: Extension to DNA“. In: *J. Chem. Theory Comput.* 11.8 (Aug. 2015), S. 3932–3945. DOI: 10.1021/acs.jctc.5b00286 (siehe S. 185 f.).

[553] Nicolas Foloppe und Alexander D. MacKerell Jr. „All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data“. In: *Journal of Computational Chemistry* 21.2 (2000), S. 86–104. ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(20000130)21:2<86::AID-JCC2>3.0.CO;2-G (siehe S. 186).

[554] S. Pronk, S. Pall, R. Schulz *et al.* „GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit“. In: *Bioinformatics* 29.7 (Feb. 2013), S. 845–854. DOI: 10.1093/bioinformatics/btt055 (siehe S. 186).

[555] Tsjerk A. Wassenaar, Kristyna Pluhackova, Rainer A. Böckmann, Siewert J. Marrink und D. Peter Tieleman. „Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models“. In: *J. Chem. Theory Comput.* 10.2 (Feb. 2014), S. 676–690. DOI: 10.1021/ct400617g (siehe S. 186).

[556] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey und Michael L. Klein. „Comparison of simple potential functions for simulating liquid water“. In: *The Journal of Chemical Physics* 79.2 (1983), S. 926. DOI: 10.1063/1.445869 (siehe S. 187).

[557] Loup Verlet. „Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules“. In: *Phys. Rev.* 159.1 (Juli 1967), S. 98–103. DOI: 10.1103/physrev.159.98 (siehe S. 187).

[558] Mark James Abraham, Teemu Murtola, Roland Schulz *et al.* „GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers“. In: *SoftwareX* 1-2 (Sep. 2015), S. 19–25. DOI: 10.1016/j.softx.2015.06.001 (siehe S. 187).

[559] Tom Darden, Lalith Perera, Leping Li und Lee Pedersen. „New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations“. In: *Structure* 7.3 (März 1999), R55–R60. DOI: 10.1016/s0969-2126(99)80033-1 (siehe S. 187).

[560] W. F. Van Gunsteren und H. J. C. Berendsen. „A Leap-frog Algorithm for Stochastic Dynamics“. In: *Molecular Simulation* 1.3 (März 1988), S. 173–185. DOI: 10.1080/08927028808080941 (siehe S. 187).

[561] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola und J. R. Haak. „Molecular dynamics with coupling to an external bath“. In: *The Journal of Chemical Physics* 81.8 (1984), S. 3684. DOI: 10.1063/1.448118 (siehe S. 187).

[562] Justin Finnerty. *Molecular dynamics meets the physical world: Thermostats and barostats.* 2011. URL: http://www.grs-sim.de/cms/upload/Carloni/Tutorials/FMCP/Thermostats_and_Barostats.pdf (besucht am 20. 08. 2016) (siehe S. 187).

[563] Giovanni Bussi, Davide Donadio und Michele Parrinello. „Canonical sampling through velocity rescaling“. In: *The Journal of Chemical Physics* 126.1 (2007), S. 014101. DOI: 10.1063/1.2408420 (siehe S. 187).

[564] M. Parrinello. „Polymorphic transitions in single crystals: A new molecular dynamics method“. In: *J. Appl. Phys.* 52.12 (1981), S. 7182. DOI: 10.1063/1.328693 (siehe S. 187).

[565] Michael Blind und Michael Blank. „Aptamer Selection Technology and Recent Advances“. In: *Molecular Therapy—Nucleic Acids* 4.1 (Jan. 2015), e223. DOI: 10.1038/mtna.2014.74 (siehe S. 204).

[566] *Random Nucleic Acid Libraries for SELEX (in vitro Selection) in Aptamer Development.* 2016. URL: http://www.trilinkbiotech.com/aptamers/aptamerlibraries.asp (siehe S. 204).

Versicherung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Hilfe eines Promotionsberaters habe ich nicht in Anspruch genommen. Weitere Personen haben von mir keine geldwerten Leistungen für Arbeiten erhalten, die nicht als solche kenntlich gemacht worden sind. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

29. November 2018 M. Sc. Rico Beier